Applied Linguistics 27/1: 1–24
doi:10.1093/applin/ami038

Language Acquisition as Rational Contingency Learning

NICK C. ELLIS

University of Michigan

This paper considers how fluent language users are rational in their language processing, their unconscious language representation systems optimally prepared for comprehension and production, how language learners are intuitive statisticians, and how acquisition can be understood as contingency learning. But there are important aspects of second language acquisition that do not appear to be rational, where input fails to become intake. The paper describes the types of situation where cognition deviates from rationality and it introduces how the apparent irrationalities of L2 acquisition result from standard phenomena of associative learning as encapsulated in the models of Rescorla and Wagner (1972) and Cheng and Holyoak (1995), which describe how cue salience, outcome importance, and the history of learning from multiple probabilistic cues affect the development of 'learned selective attention' and transfer.

This article considers how fluent language users are rational in their language processing, rational in the sense that their unconscious language representation systems are optimally prepared for comprehension and production. In this view, *language learners* are intuitive statisticians, weighing the likelihoods of interpretations and predicting which constructions are likely in the current context, and *language acquisition* is contingency learning, that is the gathering of information about the relative frequencies of form–function mappings. These arguments are well supported by the psycholinguistic evidence relating to first language. But there are important aspects of second language acquisition that do not appear to accord with this characterization, those aspects where despite massive experience of naturalistic input and usage, the system fails to become optimally tuned to represent the second language forms, their functions, and their contextualized likelihoods of occurrence. The article builds the framework for an explanation of the seeming irrationalities of L2 acquisition in terms of standard phenomena of associative learning involving 'learned selective attention.'

In order to place L1 and L2 in the context of a rational analysis of language learning, I first illustrate the problem by considering the design of word processors of a more mechanical kind than is the ultimate goal of our inquiry. Having thus set a concrete stage, I outline the process of the rational analysis of learning and memory (Anderson 1989, 1991b; Anderson and Milson 1989; Anderson and Schooler 2000). Next I describe some statistical

methods that abstract this type of information–contingency learning according to the one-way dependency statistic and the Probability Contrast Model (Cheng and Holyoak 1995), and the way that human associative learning accords the predictions of these methods (Shanks 1995). When first and second language acquisition are considered in these terms, L1 acquisition seems much more obviously rational and contingency-sensitive than does L2 acquisition. I lay the foundations for a companion article (Ellis 2006) that describes the types of situation where associative learning deviates from rationality and that argues that the apparent irrationalities of L2 acquisition result from standard phenomena of associative learning: attentional shifting in perceptual learning, latent inhibition, blocking, overshadowing, and other effects of salience, transfer, and inhibition. I describe how '*learned attention*' explains these apparently irrational effects, and how theories of animal and human associative learning include selective attention as a key component. The human learning mechanism optimizes its representations of first language from the cumulative sample of first language input. But the initial state for L2 is not a *tabula rasa*, it is a *tabula repleta*: the optimal solution for L2 is not that for L1, and L2 acquisition suffers from various types of L1 interference. Thus the shortcomings of the L2 end-state are more rational when seen through the lenses of the L1.

THE DESIGN OF AN OPTIMAL WORD PROCESSOR

Consider word processing programs you have known. You probably have strong views. Remember the one that crashed at 2 a.m. losing your only copy, the one that took ten minutes to search and replace, the one that required perverse three-letter combination commands, and the one you have now that replaces spellings, styles, and punctuations for you, whether you like it or not, and you still cannot figure how to get it to stop? Besides the obvious requirements for speed and reliability, an optimal word processing program would really know what you wanted to do next, and would present you with your next needed command, file, or figure, ready as a default but awaiting your confirmation.

There are some technical developments in this direction. One relatively new feature of several programs is that of 'Open Recent.' When a user goes to open a document, the word-processor presents a list of recently opened files to select from. If the sought after file is included in the list, the user is spared the time and effort of searching through the file hierarchy. The program is no mind-reader, it does not know the goals of the user. It simply proffers alternatives using the heuristic that the more recently a file has been opened, the more likely it is that it will be needed now. This simple rule does a surprisingly good job of putting the appropriate files on the list.

Another recent feature is that of predictive text entry. When entering options from a limited set, for example a journal name into a citation management system, we appreciate it if the system suggests the most likely

completion of our first few letters, and plumps for that choice when the uniqueness point is reached. Whole text-entry systems are being developed which work in this way (MacKay 2004; Ward and MacKay 2002). Predictive text-entry systems try to anticipate our need and complete the word stem we have begun typing, using the predictions of a statistical language model: probable pieces of text are made quick and easy to select, improbable pieces of text (low frequency words or text with spelling mistakes) are made harder to choose. Their language model learns all the time: if you use a novel word once, it is easier to write next time. These systems are the ergonomic operationalizations of the cohort theory of perception which describes how we recognize speech (Marslen-Wilson 1990). The common feature of such mechanisms is that they tune the availability of selections on the basis of past history of use: they make more likely used options more readily available.

Generally, these programs are reasonably successful. They work because events in the world tend to happen like that. Things that were likely in the past tend too to be likely in the same context today: (1) something that has been *frequently* required in the past is likely to be required now; (2) something that has been *recently* required is likely to be required now; (3) something that has been often required *in this particular context* is likely to be required now.

These principles of prediction have been formally analyzed for information need: the 'information retrieval problem,' as it has been investigated for borrowing books from libraries (Burrell 1980) or accessing files from computers, can be addressed because the statistical structure of the environment enables an optimal estimation of the odds that a particular book or file will be needed. Bayesian reasoning provides a method of reassessing probabilities in the light of new relevant information, of updating likelihoods as we gather more data. Bayes' Theorem (e.g. Bayes 1763) describes what makes an observation relevant to a particular hypothesis and it defines the maximum amount of information that can be gotten out of a given piece of evidence. Bayesian reasoning renders rationality; it binds reasoning into the physical universe (Jaynes 1996; Yudkowsky 2003).

The probability that a particular piece of information will be relevant, its 'need probability,' can thus be estimated using Bayesian evaluation procedures whereby the odds ratio for that particular piece of information is a product of its odds ratio given its particular history, that is its combined *frequency* and *recency* of occurrence, and the *current context* (Anderson 1989). Taking such factors into account, an optimal estimation of an item's need probability is possible. Using such a procedure for all items in the set, the most likely ones can then be made more available in readiness, with a cost/accuracy trade-off whereby the more items suggested, the greater the chance of a hit, but also the more false positives and the longer the list of items that have to be checked and discounted (an 'Open Recent' drop-down of thousands of items might be more inclusive, but would have no practical utility).

Current word-processing programs typically use just recency of usage in determining their suggestions. Information retrieval analysis suggests that it would be advantageous to take frequency of prior usage into account too—recently used items which also have an overall history of high frequency of usage are more likely to be needed than those of infrequent overall usage. The *iTunes*® music management program, besides providing a list of user-specified libraries, gives as a default one window showing 'Recently Played' items ordered by recency, and another window listing the 'Top 25' pieces that you have played most frequently overall. Context of use could also prove a useful guide: knowing who the particular user is and factoring that in provides a better estimate than averaging over all library or information users, as clients of *Amazon* will attest. Similarly, knowing the user is working on a particular report makes it more likely that they will need the files pertaining to figures for that report, etc. (Schooler and Anderson 1997). These are all ways of tuning a word processing package for optimal usage. As we will see, they are ways in which we human word processors are tuned for optimal operation too.

THE RATIONAL ANALYSIS OF LEARNING AND MEMORY

A major characteristic of the events and environments that are relevant to human cognition is that they are fundamentally probabilistic: as William James put it over a century ago, 'Perception is of definite and probable things' (James 1890: 82). The more information we have about this probabilistic world, then the less uncertain we are of it. The better we can predict what is going to happen next, the greater our chances of survival. Claude Shannon's development of information theory was founded on the definition that a signal conveys information to the extent that it reduces the receiver's uncertainty about the state of the world (Shannon 1948). Possession of a large body of information is good, but then you have to know how to use it in order to best predict the world. This insight has allowed, over the last thirty years or so, a sea change in our understanding of human cognition. It has stemmed in large part from researchers at Carnegie Mellon University, most notably Herb Simon, Jay McClelland, John Anderson, and Brian MacWhinney, and it concerns the ways that human cognition responds to the statistical contingencies of the world.

Rational analysis (Anderson 1990) aims to answer *why* human cognition is the way it is rather than to provide process models. Its guiding principle is that the cognitive system optimizes the adaptation of the behavior of the organism, that is that human psychological behavior can be understood in terms of the operation of a mechanism that is 'optimally adapted' to its environment in the sense that the behavior of the mechanism is as efficient as it conceivably could be—given the structure of the problem space or input–output mapping it must solve. This means that if we can find, describe, and properly characterize the problem a cognitive system is trying to solve

and find the optimal solution to this problem, then the rational analysis makes the strong prediction that the behavior of the system will correspond to this solution. And the research that has followed in testing this claim suggests that human cognition is indeed rational in this sense.

For the case of memory, for example, the optimal estimation of an item's need probability is possible. The rational analysis of memory (Anderson 1989, 1990; Anderson and Milson 1989; Schooler and Anderson 1997) considers the information need problem and the way that human memory corresponds to this needs function. Their analyses concerned the relative likelihoods of the occurrence of words in the world of discourse, and the fluency of access and relative availability of items in the mental lexicons of language users. With regard to *recency*, Schooler (1993), Schooler and Anderson (1997), and Anderson and Schooler (2000) demonstrated that there is a power (i.e. log-log linear) function relating probability of a word occurring in the headline in the *New York Times* on day *n* to how long it has been since the word previously occurred. Similarly, there are effects of *frequency*—the probability of a word occurring in, say, speech to children (from the CHILDES database), or the *New York Times*, or the electronic mail a person receives, is predicted by its past probability of occurrence.

Human cognition is sensitive to these two factors and is functionally related to them in the same way. Human memory is sensitive to *recency*: the probability of recalling an item, like the speed of its processing or recognition, is predicted by time since past occurrence. The power function relating probability of recall (or recall latency) and retention interval is known as the forgetting curve (Baddeley 1997; Ebbinghaus 1885). Indeed our rate of forgetting perfectly reflects the decreasing power function of time with which information becomes redundant in the environment (Wixted and Ebbesen 1991). The forgetting curve applies to linguistic constructions and other contents of memory alike. Human learning is sensitive to *frequency*: the more times a stimulus is encountered, the faster and more accurately it is processed. The power function relating accuracy and prior occurrence frequency is known as the power law of learning (Anderson 1982; DeKeyser 2001; Ellis and Schmidt 1998; Newell 1990; Newell and Rosenbloom 1981). This describes the relationships between practice and performance in the acquisition of a wide range of cognitive contents and skills, linguistic and non-linguistic alike, whereby the effects of practice are greatest at early stages of learning but they eventually reach asymptote, as is evident for example in the practice gains that result from increasing frequency of experience in reading accuracy and rate (Plaut *et al*. 1996), picture naming (Oldfield and Wingfield 1965), typing, speaking, and signing (Kirsner 1994), and morphological processing (DeKeyser 2001; Ellis and Schmidt 1998). Human processing is sensitive to both these independent variables of *recency* and *frequency* additively: processing accuracy (or fluency) follows a joint power function of retention interval and *frequency* (Schooler and Anderson 1997).

So much for frequency and recency, but what of *context*? Schooler additionally showed that these effects are qualified by immediate context: a particular word is more likely to occur when other words that have co-occurred with it in the past are present. Schooler used the example that a headline one day mentioned Qaddafi and Libya, and sure enough a headline the next day that mentioned Qaddafi also mentioned Libya. I am sure you could give a similar collocation pair from your experience of the news this very week, given the ubiquity of collocations and the idiom principle (Biber *et al*. 2004; Biber *et al*. 1998; Schmitt 2004). Schooler collected likelihood ratio measures of association between various words in order to assess the effect of this local context factor on memory and processing. As already described, in both the child language and the *New York Times* databases, a word was more likely to occur if it had occurred previously, but additionally, a word was more likely to occur if a string associate of it occurred, and these effects were additive in the way predicted by Bayesian probability (Bayes 1763; Yudkowsky 2003). Such effects of context affect human language processing too: Schooler showed that word fragment completion was harder for words shown alone out of context (SEA____? or FAC___?) than it was for the second word of a strong context collocation (as in HERMETICALLY-SEA____? or LANGUAGE-FAC___?), a particular example of the more general phenomenon of priming whereby lexical recognition is faster when primed by an appropriate semantic or contextual constraint (Hodgson 1991; Williams 1996). We process collocates faster and we are more inclined therefore to identify them as a unit. These processing effects are crucial in the interpretation of meaning—it is thus that an idiomatic meaning can overtake a literal interpretation, and that familiar constructions can be perceived as wholes. The effects are crucial in our production of language, too (Bybee and Hopper 2001). Lexical cognition, our learning, memory, and processing of words, is rationally tuned to the likelihoods of occurrence of words as they behave in the world. What a facility this ever-updated model of the world provides: as information is lost in the world, so it is lost in our minds; as it becomes relevant in the world, so our minds make it available to us.

But prediction is never 100 percent accurate; there is always some error, a trade-off between false positives and misses. Could the costs of a wrong prediction and subsequent backtracking not outweigh the computational benefits of correct predictions? The considerable investment of the computer industry over the last twenty years into ways of achieving normative prediction in its processors (Lee and Smith 1984) suggests otherwise. A significant design improvement of the Pentium-4 chip was 'branch prediction,' a method that attempts to predict what will need to be fed to the processor next. The processor calculates, and branch prediction attempts to help, by guessing what the processor will need next. The predictive model is updated online by a program for 'advanced dynamic execution' which keeps track of what worked in the branch prediction and what did not,

this helping overall to reduce branch mis-prediction by about 33 percent. The Macintosh G5 processor continues development of the use of branch prediction to the point where the algorithm anticipates which instruction will occur next in a sequence, with speculative operation causing that instruction to be executed. If the prediction is correct, the processor works more efficiently, since the speculative operation has executed an instruction before it is required. If the prediction is incorrect, the cost is that the processor must clear the unneeded instruction and associated data, resulting in an empty space called a pipeline bubble, a performance killer as the processor marks time waiting for the next instruction to present itself. IBM and Apple claim the G5 can predict branch processes with an accuracy of up to 95 percent. So, while humankind still struggles to build rationality into the design of our information processing machines, our computers, devices, and software (Norman 1988, 1993), our minds themselves are richly endowed with implicit optimal processing and memory. This is a neurobiological heritage. The wide gamut of animal conditioning phenomena are explainable in terms of information gain and rational analysis (Gallistel 2003), even the lowly mollusk *Hermissenda* evidences contingency learning (Farley 1987). There is a wonderful irony in the observation that current research into branch prediction algorithms for computers is looking to neural network methods based on simple perception models in order to improve performance accuracy (Jiménez and Lin 2002).

In their classic review of human learning, Peterson and Beach (1967) identified that human learning is to all intents and purposes perfectly calibrated with normative statistical measures of contingency like r, χ^2 and ΔP (which are explained in detail on pp. 10–12 in the 'Statistical Learning Methods' section of this paper), and that probability theory and statistics provided a firm basis for psychological models that integrate and account for human performance in a wide range of inferential tasks. They entitled their paper 'Man as an intuitive statistician.'

PROBABILISTIC LANGUAGE PROCESSING

Fluent language processing, too, is exquisitely sensitive to frequency of usage. In Ellis (2002a), I reviewed evidence that language performance is tuned to input frequency at all sizes of grain: phonology and phonotactics, reading, spelling, lexis, syntax and morphosyntax, grammaticality, formulaic language, language comprehension, and sentence production. There is good evidence that human implicit cognition, acquired over natural ecological sampling as natural frequencies on an observation-by-observation basis, is rational in this sense (Anderson 1990, 1991a, 1991b; Gigerenzer and Hoffrage 1995; Sedlmeier and Betsc 2002; Sedlmeier and Gigerenzer 2001). Psycholinguistics is the testament of rational language processing and the usage model. The words that we are likely to hear next, their most likely senses, the linguistic constructions we are most likely to utter next,

the syllables we are likely to hear next, the graphemes we are likely to read next, and the rest of what is coming next across all levels of language representation, are made more readily available to us by our language processing systems. Not only do we know the constructions that are most likely to be of overall relevance (i.e. first-order probabilities of occurrence), but we also predict the ones that are going to pertain in any particular context (sequential dependencies), and the particular interpretations of cues that are most likely to be correct (contingency statistics). These predictions are usually rational and normative in that they accurately represent the statistical covariation between events. In these ways, language learners are intuitive statisticians; they acquire knowledge of the contingency relationships of one-way dependencies and they combine information from multiple cues.

Consider, for example, that while you are conscious of words in your attentional focus, you certainly did not consciously label the word 'focus' just now as a noun; yet this sentence would be incomprehensible if your unconscious language analyzers did not treat 'focus' as a noun rather than as a verb or an adjective. Nor, on reading 'focus,' were you aware of its nine alternative meanings or of their rankings in overall likelihood, or of their rankings in this particular context, rather than in different sentences where you would instantly bring a different meaning to mind. A wealth of psycholinguistic evidence suggests that this information is available unconsciously for a few tenths of a second before your brain plumps for the most appropriate one in this context. Most words have multiple meanings, but only one at a time becomes conscious. This is a fundamental fact about consciousness (Baars 1988, 1997). In these ways, our unconscious language mechanisms present up to consciousness the constructions that are most likely to be relevant next. Their offerings are usually appropriate, but consciousness can decline if it has reason to think better. In sum, there is good reason to view the unconscious mechanisms of fluent language users as operating as optimal word processors. They are adaptively probability-tuned to predict the linguistic constructions that are most likely to be relevant in the ongoing discourse context.

The evidence of rational language processing implies that language learning too is an intuitive statistical learning problem, one that involves the associative learning of representations that reflect the probabilities of occurrence of form–function mappings. Learners have to FIGURE language out: their task is, in essence, to learn the probability distribution P(*interpretation*|*cue, context*), the probability of an interpretation given a formal cue in a particular context, a mapping from form to meaning conditioned by context (Manning 2003). In order to achieve optimal processing, acquisition mechanisms must have gathered the normative evidence that is the necessary foundation for rationality. To accurately predict what is going to happen next, we require a representative sample of experience of similar circumstances upon which to base our judgments,

and the best sample we could possibly have is the totality of our linguistic experience to date. Usage-based theories hold that an individual's linguistic competence emerges from the collaboration of the memories of all of the utterances in their entire history of language input and use. The systematicities of language competence, at all levels of analysis from phonology, through syntax, to discourse, emerge from learners' lifetime analysis of the distributional characteristics of the language input and their usage. It is these ideas that underpin the last thirty years of investigations of cognition using connectionist and statistical models (Christiansen and Chater 2001; Elman *et al*. 1996; Rumelhart and McClelland 1986), the competition model of language learning and processing (Bates and MacWhinney 1987; MacWhinney 1987, 1997), usage-based models of acquisition (Barlow and Kemmer 2000; Langacker 1987, 2000; Tomasello 1998, 2003), the recent emphasis on frequency in language acquisition and processing (Bod *et al*. 2003; Bybee and Hopper 2001; Ellis 2002a, 2002b; Jurafsky 2002; Jurafsky and Martin 2000) and in NLP (Jurafsky and Martin 2000; Manning and Schuetze 1999), and proper empirical investigations of the structure of language by means of corpus analysis (Biber *et al*. 1998; Biber *et al*. 1999; Sampson 2001; Sinclair 1991).

The proposal, then, is that L1 acquisition and fluent processing are as rational as other aspects of human learning and memory, and that they can be understood according to standard principles of associative learning. What are the learning mechanisms and mental algorithms that compute these norms? In the next section, I review evidence that the appropriate normative theory of the learning of these associations is contingency theory. But your consideration of these associative learning accounts of first language acquisition should reflect as well those aspects of second language acquisition that do not appear to be rational—the fragile features of language that L2 learners fail to acquire despite thousands of occurrences in their input, the cases where input fails to become intake. Does this mean that L2 acquisition cannot be understood according to the general principles of associative learning that underpin other aspects of human cognition, that L2 acquisition is fundamentally *irrational*? Or, paradoxically, does associative learning theory explain these limitations too?

STATISTICAL LEARNING METHODS

First order probability: tallying frequencies

We are more likely to perceive things that are more likely to occur. The power law of learning describes how the resting levels of detectors for words, letters, and other linguistic constructions are set according to their overall frequency of usage so that less sensory evidence is needed for the recognition of high frequency stimuli than for low frequency stimuli. Each time we process a stimulus, there is a practice increment whereby the

resting strength of its detector is incremented slightly, resulting in priming and a slight reduction in processing time the next time this stimulus is encountered. Ellis (2002a) summarizes evidence (1) that we have implicit rank order information for the relative frequencies of letters, bigrams, words, and the wide range of other linguistic constructions; (2) that neurobiological learning processes underpin this tallying of occurrence, and (3) that the strengthening function that relates frequency and resting state is not linear but instead follows the power law of learning with the effects of practice being greatest at early stages of learning but eventually reaching asymptote: In these ways the perceptual and motor systems become tuned to the relative frequencies of individual constructions (Sedlmeier and Betsc 2002).

Contingency: ΔP

Classical conditioning involves a cue (a to-be-conditioned stimulus, CS, for example, a bell) being temporally paired with an outcome (an unconditioned stimulus, US, for example, food), with, after several such pairings, the animal emitting a conditioned response (CR, salivation) on encountering the cue alone. The initial interpretation of this phenomenon was that it was the temporal pairing of the CS and the US that was important for learning to take place. However, Rescorla (1968) showed that if one removed the contingency between the CS and the US, preserving the temporal pairing between CS and US but adding additional trials where the US appeared on its own, then animals did not develop a conditioned response to the CS. This result was a milestone in the development of learning theory because it implied that it was contingency, not temporal pairing, that generated conditioned responding. It was as if in Rescorla's experiment *the rats were acting as scientists*, picking up on cues in the environment if they had value in predicting what was going to happen next. The rats were behaving rationally. Contingency, and its associated aspects of predictive value, information gain, and statistical association, have been at the core of learning theory ever since.

Every social scientist is used to the inferential statistical methods that are used to test association between two variables. If the variables are continuous we use correlational methods, like *r* or *rho*, to determine the degree to which we can predict position on one dimension from knowledge of position on the other. If the variables are categorical, then we lay the data out in a contingency table such as Table 1, we count the number of observations that fall into each of the cells, and we use nonparametric methods such as χ^2, lambda, or one of the other possibilities offered by our stats package following a crosstabulation analysis, to look for a contingency between the rows and the columns.

But χ^2 is a measure of the two-way dependency between a pair of events. The directional association between a cue and an outcome, as illustrated in

Table 1: A contingency table showing the four possible combinations of events showing the presence or absence of a target Cue and an Outcome

	Outcome	No outcome
Cue	a	b
No cue	c	d

Notes. a, b, c, d represent frequencies, so, for example, a is the frequency of conjunctions of the cue and the outcome, and c is the number of times the outcome occurred without the cue.

Table 1, is better measured using the one-way dependency statistic ΔP (Allan 1980):

$$\begin{aligned}\Delta P &= P(O|C) - P(O|{-}C)\\ &= a/(a+b) - c/(c+d)\\ &= (ad - bc)/[(a+b)(c+d)]\end{aligned}$$

ΔP is the probability of the outcome given the cue P(O|C) minus the probability of the outcome in the absence of the cue P(O|–C). When these are the same, when the outcome is just as likely when the cue is present as when it is not, there is no covariation between the two events and $\Delta P = 0$. ΔP approaches 1.0 as the presence of the cue increases the likelihood of the outcome and approaches –1.0 as the cue decreases the chance of the outcome—a negative association.

The last thirty years have evidenced many psychological investigations into human sensitivity to the contingency between cues and outcomes in laboratory tasks involving estimation, for example, of the influence of pressing a telegraph key on the chance of a light coming on, or the degree to which a symptom is indicative of a disease in medical diagnosis. Many of these experiments are assembled in chapter 2 of the excellent Shanks (1995). Shanks' conclusion is that humans' associative judgments in such situations are unbiased at asymptote and that, when given sufficient exposure to a relationship, people's judgments match quite closely the contingency specified by normative ΔP theory. Biases may occur prior to asymptote, with judgments only slowly regressing towards the predicted values, because a reasonably large sample of events is required by associative learning mechanisms such as the delta rule to compute the contingency, but at asymptote the contingency judgments are, to all intents and purposes, normative. For one example, consider the study of Wasserman *et al*. (1993) who had participants judge the extent to which their pressing a telegraph key caused a light to flash in twenty-five different problems crossing every combination of settings of P(O|C) and P(O|−C) at 0.0, 0.25, 0.5, 0.75, and 1.0. The participants' judgments of contingency explained 96.7 percent

of the variance of the actual values—an impressive degree of sensitivity to contingency.

It is this sensitivity that underpins our ability to rank order the frequencies of occurrence of bigrams (Hasher and Chromiak 1977), to tune our processing system to those other sequential dependencies of language, and to recognize the interpretations (outcomes) that are most relevant to particular formal constructions (cues), with no cue being totally unambiguous (Ellis 2002a, 2003). Learning language can thus be viewed as a statistical process in that it requires the learner to acquire a set of likelihood-weighted associations between constructions and their functional/semantic interpretations.

These processes of rational learning over natural ecological sampling on an observation-by-observation basis (Anderson 1990, 1991a, 1991b; Gigerenzer and Hoffrage 1995; Sedlmeier and Betsc 2002; Sedlmeier and Gigerenzer 2001) have been intensively investigated in the last thirty years of work within the Connectionist tradition. Connectionist models have been successful in simulating a wide range of human inductive phenomena in the perception and classification of linguistic and non-linguistic domains alike (Christiansen and Chater 2001; Ellis 1998; Elman *et al*. 1996; Rumelhart and McClelland 1986). These systems learn by being exposed to input cues, by making a prediction of outcome, and on the realization of whether their prediction was correct or not, adjusting the weights of the connections between their processing units (their synaptic strengths) so that their prediction of outcome would be more accurate if faced with the same situation again. The standard and simplest connectionist learning algorithm for the incremental tuning of weights using backpropagation of error is the delta rule (Widrow and Hoff 1960). The delta rule can be shown to compute ΔP at asymptote when other background cues are constant (Chapman and Robbins 1990). Thus connectionist and human learners match quite closely the contingencies specified by normative ΔP theory.

For the analysis of first and second language acquisition in these *statistical connectionist* terms, for the measurement and simulation of these phenomena, and for analysis of their effects in sentence processing, we can again profitably look to Carnegie Mellon University: The Competition Model is the most extensive single account of these statistical phenomena as they underpin the emergence of language (Bates and MacWhinney 1987; MacWhinney 1987, 1997, 2001a, 2001b).

Multiple cues to interpretation: the probabilistic contrast model

Normative ΔP theory describes associative learning where learners have to acquire the relationship between a cue and an outcome and where the cue is the only obvious causal feature present. In such situations, contingency is easy to specify and human learning is shown to be rational in that it accords

with the normative ΔP rule. However, it is rarely, if ever, the case that predictive cues appear in isolation, and most utterances, like most other stimuli, present the learner with a set of cues which co-occur with one another, with the learner's task being to determine the ones that are truly predictive. In such cases of multiple cues to interpretation, then, the predictions of normative analysis using the ΔP rule are muddied by *selection effects*: learners selectively choose between potential causal factors. Thus, in some circumstances, the cue may be selected for association with an interpretation whilst in other circumstances it may not, depending on the presence and status of other cues.

Statisticians do not have an agreed procedure for specifying the contingency between events C and O when the background varies. However, the psychologists Cheng and Hoyoak (1995) and Cheng and Novick (1990) have proposed an extended version of contingency theory, which they termed the Probabilistic Contrast Model (PCM), as a descriptive account of the use of statistical regularity in human causal induction. The model, which applies to events describable by discrete variables, assumes that potential causes are evaluated by contrasts computed over a 'focal set.' The focal set for a contrast is a contextually-determined set of events that the reasoner *selects* to use as input to the computation of that contrast. The focal set consists of all trials on which the target cue is present as well as all those trials that are identical to the target present trials except for the absence of the target. Thus it is often not the universal set of events, contrary to what had been assumed by previous contingency theories in psychology, and hence the results of this reasoning appear irrational when measured unconditionally against the entire input across all learning trials. Yet despite this, PCM theory does measure up well against the logic of classical scientific method. Thus one implication of Cheng and Holyoak's 'Adaptive systems as intuitive statisticians' paper is that people are 'intuitive scientists,' a sentiment expressed earlier in classic attribution theory by Kelley (1967) and in personal construct theory by George Kelly: 'For Kelly, all men can be said to be ''scientists'' in the sense that they have theories about their universe (not as systematic or sophisticated as the theories of professional scientists but theories nevertheless) and on the basis of these theories they have particular hypotheses (expectations) which are fulfilled or not fulfilled, and in the light of the outcome of their ''experiments'' their views are modified. Thus the model man of Personal Construct Theory is ''man the scientist''' (Bannister and Fransella 1986: 362).

So what are the algorithms and outcomes of PCM that lead to this claim that people are intuitively scientific, both in their methods and in their optimality?

First the algorithm. In order to determine whether C is a valid cue of O, the PCM procedure is to calculate P(O|C) for the target cue C across trials on which C occurs, and to calculate P(O|–C) across trials that are identical to the C trials with the exception that C is absent. Thus ΔP is calculated not across

all trials, but across a subset of trials (the focal set) in which the background effects are kept constant. Consider, for an example of this in language learning, sentence processing where comprehenders must assign nouns to linguistic roles such as actor, patient, and recipient, the appropriate assignments being predicted by various cues of varying reliability. These cues include word order, noun animacy, and case inflection. As in the example sentence *the dog chased him*, the actor usually precedes the patient, the actor is usually animate, and case inflection can differentiate between actor (*he*) and patient (*him*). These cues may or may not be present in every sentence, and cues may at times conflict or 'compete' with each other, as in the example *the television smashed the dog* (MacWhinney *et al*. 1984). As we will see in the next section, learners do not assess the predictive power of these cues over all of the sentences they are exposed to, many of them including several such cues, for example *the ball was chased by the cat*, *the televisions smash the dogs*, etc. Instead, they may try to determine the success of outcomes of assignments based on one cue at a time, using relevant focal sets where the other cues are kept constant, for example *the dog chased the cat* vs. *the cat chased the dog*. The parallel with the experimental method is clear: classically, the only difference between the experimental and control conditions is the independent variable of concern, with all other potential independent variables being held constant. By these means, each experiment usually focuses upon just one potential cue at a time while the rest go ignored and unconsidered (Ellis, in press).

Secondly, the outcomes. Cheng and Holyoak (1995) argued not only that the PCM is the appropriate normative theory for causal or associative relationships when the background is variable, but also that human behavior is closely matched to it. They and Shanks (1995: ch. 2) provide a range of results of video-game and medical-diagnosis-based multiple cue contingency judgment tasks which appear irrational when measured against ΔP theory applied to the whole learning set, but which are much better accommodated by the PCM extension.

In contrast to my prior claim about the rationality of human implicit cognition acquired over natural ecological sampling as natural frequencies on an observation-by-observation basis (Anderson 1990, 1991a, 1991b; Gigerenzer and Hoffrage 1995; Sedlmeier and Betsc 2002; Sedlmeier and Gigerenzer 2001), there are various demonstrations from Kahneman and Tversky (1972) onwards that human conscious inference deviates from Bayesian inference. The way that human everyday statistical/scientific reasoning is not rational is that it tends to neglect the base rates, the prior research findings. When people approach a problem where there is some evidence X indicating that hypothesis A might hold true, they tend to judge A's likelihood solely by how well the current evidence X seems to match A, without taking into account the prior frequency or probability of A (Tversky and Kahneman 1982).

So humans intuitively make their judgments of contingency between potential cues and outcomes according to the probability contrast model which approximates the scientific method. And like scientists, they too see the world with focal vision, through lenses of selective attention: 'Man tries to make for himself in the fashion that suits him best a simplified and intelligible picture of the world: he then tries to some extent to substitute this cosmos of his for the world of experience, and thus to overcome it. This is what the painter, the poet, the speculative philosopher and the natural scientist do each in his own fashion' (Albert Einstein address delivered to the Physical Society of Berlin in 1918).

PCM and language learning

Is the sequence of cue acquisition in language learning also as the PCM would predict? It appears so. Experiments using miniature artificial languages have shown that, in the initial stages of acquisition, learners tend to focus on only one cue at a time (Blackwell 1995; MacWhinney and Bates 1989; Matessa and Anderson 2000; McDonald 1986; McDonald and MacWhinney 1991). For example, when cues for determining the agent in sentences include word order, noun animacy and agreement of noun and verb, learners typically decide to focus attention on only one of these as the predictor of interpretation. MacWhinney *et al.* (1985) demonstrated that the cue that children first focus upon is that which has the highest overall validity as measured by its availability (its frequency or probability of occurrence) times its reliability (its probability of correctly indicating the interpretation, broadly equivalent to its ΔP). The effect is that a cue with high availability but low reliability may initially be used over a cue that is of lower availability, even though it is in fact more reliable. Learners focus on one cue alone to begin with. Later on, after having tracked the use of this first cue, they will add a second cue to the mix and begin to use the two in combination, and, as development proceeds, so additional cues may be added if they significantly help reduce errors of understanding, as measured by the statistic 'conflict validity' which relates to how the cue affords extra predictive accuracy when its interpretation conflicts with that of a co-occurring cue. This variable-by-variable incremental sequence is as predicted by the probability contrast model.

ASSOCIATIVE LEARNING AND ATTENTION

What at first sight was temporal association proved to be contingency learning. And what at first sight was mere tallying, with the learner weighing all cues equally, proved to be subject to selection effects. The driving forces of language learning, then, are frequency, conditioned by contingency, conditioned by selection. But there is still more. There are effects of attention beyond those of selection for scrutiny in the PCM account

of the assessment of contingencies. These are effects of salience, and overshadowing, and learned attention, and not for the first time, these notions have a long and distinguished history in associative learning theory.

Selective attention is not the preserve of scientists, nor of creative people, nor even of every human body else; it is a key aspect of the behavior of all organisms. Reynolds (1961) trained two pigeons to peck the red key with a white triangle for food reward. Pecking the green key with the white circle did not yield reward. When he tested each of the four stimulus components in isolation, he found that both birds were strongly conditioned, but *to different aspects* of their identical training experience: one bird was conditioned to the white triangle, while the other was conditioned to the red background. Each bird responded to one of the two elements composing the positive stimulus to the exclusion of the other element. It was this that led Reynolds to introduce the notion of selective attention into the learning literature and it has remained a core component of conditioning theory ever since.

Experimental investigations of selective attention between multiple cues illustrate the robust phenomenon of *overshadowing*. In such experiments, two cues, C1 and C2, are always presented together during training and they jointly predict an outcome. In the test-phase, the strength of conditioning to C1 and C2 presented individually are measured. The typical outcome is that the strength of conditioning to each cue depends on their relative intensity. If C1 is a dim light and C2 a bright light then, after conditioning to the C1–C2 combination, the learned response to the bright light is very strong while the dim light alone produces little or no reaction (Kamin 1969). Wagner *et al.* (1968) showed that when one cue is more reliably informative of outcome like this, it is the only one to which the conditioned response develops. It does not develop to the other less reliably informative CSs, even though they are frequently paired with the US. There are two important aspects of these overshadowing results. The first relates to the salience that causes the learned cue to be learned. The second relates to expectancy, habituation, and surprise in the overshadowing process.

The general perceived strength of stimuli is commonly referred to as their *salience*. Although it might in part be related to the physically measurable intensity of stimuli, salience refers to the intensity of the subjective experience of stimuli, not of the objective intensity of the stimuli themselves. Salience, as subjective experience, varies between individuals, and, more importantly, between species. Kamin (1969) interpreted the phenomenon of overshadowing as implying that in such situations the animal 'expected' the outcome because of the cue provided by the more salient CS, that is, it was not 'surprised' by it, and thus pairing it with an additional cue did not produce a conditioned response. Rescorla and Wagner (1972) presented

a formal model of conditioning which expresses the capacity any cue (CS) has to become associated with an outcome (US) at any given time. This associative strength of the US to the CS is referred to by the letter V and the change in this strength which occurs on each trial of conditioning is called dV. The more a CS is associated with a US, the less additional association the US can induce. This informal explanation of the role of US surprise and of CS (and US) salience in the process of conditioning can be stated as follows:

$$dV = ab(L - V)$$

where a is the salience of the US, b is the salience of the CS, and L is the amount of processing given to a completely unpredicted US. So the salience of the cue and the importance of the outcome are essential factors in any associative learning. The Rescorla–Wagner model pulled together the findings of hundreds of experiments each designed with an empirical rigour unsurpassed outside animal learning research. Its generality of relevance makes it arguably the most influential formula in the history of conditioning theory.

It might come as a surprise to see such terms as *selective attention*, *salience*, *expectation*, and *surprise* being bandied about as explanations of animal learning. But the experimental findings simply could not be explained without these concepts, and even though researchers like Kamin (1968), coming from a behaviorist background which eschewed such anthropomorphic concepts along with all other speculations of what went on inside the black box, felt obliged to keep scare quotes around these terms, such notions rapidly became key elements of associative learning theory (Mackintosh 1975; Pearce and Bouton 2000), and their investigation became standard fare in undergraduate animal practical classes: indeed you yourself could further explore these phenomena now on your PC if you wished, using 'Sniffy the Virtual Rat' (Krames *et al*. 1997).

However, I know associative learning theory is not the usual fare of *Applied Linguistics*. Pigeons' lack of pidgin, and Sniffy's apparent inability at any form of language, might well have you turning your nose up at all this animal work. 'Too much learning, too little language,' you may well be thinking, however illustrative these animal experiments are of the generality of these associative learning phenomena. But bear with me, for herein, I believe, lie important insights into first and second language acquisition both, not only for the difficulties and ordering of acquisition of different grammatical constructions in L1, but also perhaps for the biggest conundrum of all, the apparent irrationality of the shortcomings of L2 acquisition and of fossilization. In drawing a close on this first article, I will summarize the problem and briefly gather the components that have been introduced here that will, I believe, provide some solutions when applied in more detail in the companion piece to follow.

APPARENT DEVIATIONS FROM RATIONALITY IN L2 ACQUISITION

A founding observation, from the very beginnings of applied linguistics, is that although learners are surrounded by language, not all of it 'goes in': this is Corder's distinction between input, the available target language, and intake, that subset of input that actually gets in and which the learner utilizes in some way (Corder 1967). What are the fragile aspects of language to which second language learners commonly prove impervious, where input fails to become intake?

Schumann (1978), on the basis of his analysis of the ESL of Alberto, a 33-year-old Costa Rican polisher, likened the second language acquisition process to one of pidginization. Alberto's spontaneous conversations and elicited language over a ten month period evidenced the lack of a variety of grammatical constructions including negative placement, question inversion, suppliance of grammatical morphemes such as possessive '-s' forms, regular past tense, and progressive '-ing,' and most auxiliaries apart from 'can.' He concluded that 'In general Alberto can be characterized as using a reduced and simplified form of English' (Schumann 1978: 65), resembling pidgins. Indeed, Schumann's analysis showed that the more that pidgin speakers 'use each morpheme, the higher the percentage of correct use for Alberto' (Schumann 1978: 187). In pidgins, there is usually only a minimal pronoun system without gender or case, there is an absence of agreement markers for number or negation, there is no inflectional morphology; only the bare essentials necessary for communication are present.

Schmidt's (1984) case study of naturalistic language learner, Wes, showed him to be very fluent, with high levels of strategic competence, but low levels of grammatical accuracy: 'using 90 percent correct in obligatory contexts as the criterion for acquisition, none of the grammatical morphemes counted has changed from unacquired to acquired status over a five year period' (Schmidt 1984: 5). At a recent AAAL conference where I quoted that figure, Schmidt affirmed that the same could still be said of Wes' interlanguage over what is now a twenty-five year period.

Larger and more representative samples can be found in the ESF cross-linguistic and longitudinal research project (Perdue 1993) which examined how 40 adult learners picked up the language of their social environment by everyday communication. Analysis of the interlanguage of these L2 learners resulted in its being described as the 'Basic Variety.' All learners, independent of source language and target language, developed and used it, with about one-third of them fossilizing at this level in that although they learned more words, they did not further complexify their utterances in respects of morphology or syntax. In this Basic Variety, most lexical items stem from the target language, but they are uninflected. 'There is no functional morphology. By far most lexical items correspond to nouns, verbs and adverbs; closed-class items, in particular determiners, subordinating elements,

and prepositions, are rare, if present at all. . . . Note that there is no functional inflection whatsoever: no tense, no aspect, no mood, no agreement, no casemarking, no gender assignment; nor are there, for example, any expletive elements' (Klein 1998: 544–5).

These morphemes abound in the input, but they are simply not picked up by learners. What are the factors, then, that modulate the effects of simple availability?

SOME SYNTHESIS AND A PROMISSORY NOTE

In general, the linguistic forms that L2 learners fail to adopt and to use routinely thereafter in their second language processing are those which, however available as a result of frequency, recency, or context, fall short of intake because of one of the associative learning factors that have been described here:

1 They are unreliable predictors of outcome, with contingency statistics such as ΔP falling far short of 1.0 (Bates and MacWhinney 1987; MacWhinney 1987, 2001a).
2 They fail to be attended in the PCM selection process because of low cue salience (Andersen 1984, 1990).
3 They fail to be attended in the PCM selection process because of low importance of functional outcome in the overall interpretation of the message.
4 They fail to be attended because they are redundant in the immediate understanding of an utterance, being overshadowed or blocked by higher salience cues which have previously been selected. As summarized in Rescorla-Wagner, the more a cue is associated with an interpretation, the less additional association that cue can induce, and obversely, the more predicted the interpretation from context and other cues, the less the additional association from an associated cue on this trial.
5 They are ignored because the multitude of form ↔ meaning contingencies acquired from input and usage conspire to tune the ways in which we selectively attend in our processing of language. What emerges, as detailed in my companion article in the next issue, is that L1 experience of form → meaning contingencies affects the cues and dimensions that an L2 learner's language input systems can best distinguish (perceptual learning), and L1 experience of meaning → form contingencies affects the way an L2 learner routinely expresses their meanings in language ('thinking for speaking,' (Slobin 1996)).

Factors such as ΔP, and salience, and outcome importance are going to affect L1 acquisition too, and so play a role in L1 acquisition and L2 acquisition, although, if learning/computational resources in older brains are generally reduced, they may result in more pronounced effects in older learners of

a second language. But redundancy, blocking, overshadowing, L1 content interference, and L1 perceptual tuning could all have a differential, more marked role on L2 acquisition, thus helping to provide non age-invoked biological explanations for why L2 acquisition stops short while first language acquisition does not.

Factors 2–5 all concern attention, in one way or another. Schmidt opens the recent collection on 'Cognition and Second Language Instruction' (Robinson 2001) with the essential claim that 'the concept of attention is necessary in order to understand virtually every aspect of second language acquisition (L2 acquisition), including the development of interlanguages over time, variation within IL at particular points in time, the development of L2 fluency, the role of individual differences such as motivation, aptitude, and learning strategies in L2 learning, and the ways in which interaction, negotiation for meaning, and all forms of instruction contribute to language learning' (Schmidt 2001: 3). In my companion article (Ellis 2006) to appear in the next issue of this journal, I take Schmidt's lead concerning the role of attention in IL development, illuminating the language acquisition phenomena and explaining in more detail how attention becomes tuned by the psychological learning processes outlined above. In a third paper (Ellis 2005), I consider the implications of this theory for instruction and why the shortcomings of L2 acquisition are best remedied using techniques of attentional refocus and explicit learning.

Final version received May 2005

ACKNOWLEDGEMENTS

I thank Gabi Kasper, Martin Bygate, Robert DeKeyser, and anonymous *Applied Linguistics* Readers for insightful and constructive advice on a prior version of this paper.

REFERENCES

Allan, L. G. 1980. 'A note on measurement of contingency between two binary variables in judgment tasks,' *Bulletin of the Psychonomic Society* 15: 147–9.

Andersen, R. W. 1990. 'Models, processes, principles and strategies: second language acquisition inside and outside of the classroom' in B. Van Patten and J. Lee (eds): *Second Language Acquisition—Foreign Language Learning*. Clevedon: Multilingual Matters, pp. 45–68.

Andersen, R. W. (ed.). 1984. *Second Language: A Crosslinguistic Perspective*. Rowley, MA: Newbury House.

Anderson, J. R. 1982. 'Acquisition of cognitive skill,' *Psychological Review* 89/4: 369–406.

Anderson, J. R. 1989. 'A rational analysis of human memory' in H. L. I. Roediger and F. I. M. Craik (eds): *Varieties of Memory and Consciousness: Essays in honour of Endel Tulving*. pp. 195–210.

Anderson, J. R. 1990. *The Adaptive Character of Thought*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Anderson, J. R. 1991a. 'The adaptive nature of human categorization,' *Psychological Review* 98/3: 409–29.

Anderson, J. R. 1991b. 'Is human cognition adaptive?' *Behavioral & Brain Sciences* 14/3: 471–517.

Anderson, J. R. and **R. Milson.** 1989. 'Human memory: An adaptive perspective,' *Psychological Review* 96/4: 703–19.

Anderson, J. R. and **L. J. Schooler.** 2000. 'The adaptive nature of memory' in E. Tulving and F. I. M. Craik (eds): *The Oxford Handbook of Memory*. London: Oxford University Press, pp. 557–70.

Baars, B. J. 1988. *A Cognitive Theory of Consciousness*. Cambridge: Cambridge University Press.

Baars, B. J. 1997. *In the Theater of Consciousness: The Workspace of the Mind*. Oxford: Oxford University Press.

Baddeley, A. D. 1997. *Human Memory: Theory and Practice*, Rev edn. Hove: Psychology Press.

Bannister, D. and **F. Fransella.** 1986. *Inquiring Man: The Psychology of Personal Constructs* (3rd edn). London: Croom Helm.

Barlow, M. and **S. Kemmer.** (eds). 2000. *Usage Based Models of Language*. Stanford, CA: CSLI Publications.

Bates, E. and **B. MacWhinney.** 1987. 'Competition, variation, and language learning' in B. MacWhinney (ed.): *Mechanisms of Language Acquisition*. pp. 157–93.

Bayes, T. 1763. 'An essay towards solving a problem in the doctrine of chances,' *Philosophical Transactions of the Royal Society of London* 53: 370–418.

Biber, D., S. Conrad, and **V. Cortes.** 2004. ''"If you look at . . ."': Lexical bundles in university teaching and textbooks,' *Applied Linguistics* 25, 371–405.

Biber, D., S. Conrad, and **R. Reppen.** 1998. *Corpus Linguistics: Investigating Language Structure and Use*. New York: Cambridge University Press.

Biber, D., S. Johansson, G. Leech, S. Conrad, and **E. Finegan,** 1999. *Longman Grammar of Spoken and Written English*. Harlow, UK: Pearson Education.

Blackwell, A. 1995. Artificial languages, virtual brains. Unpublished doctoral dissertation, University of California at San Diego.

Bod, R., J. Hay, and **S. Jannedy.** (eds). 2003. *Probabilistic Linguistics*. Cambridge, MA: MIT Press.

Burrell, Q. L. 1980. 'A simple stochastic model for library loans,' *Journal of Documentation* 36: 115–32.

Bybee, J. and **P. Hopper.** (eds). 2001. *Frequency and the Emergence of Linguistic Structure*. Amsterdam: Benjamins.

Chapman, G. B. and **S. J. Robbins.** 1990. 'Cue interaction in human contingency judgment,' *Memory & Cognition* 18: 537–45.

Cheng, P. W. and **K. J. Holyoak.** 1995. 'Adaptive systems as intuitive statisticians: Causality, contingency, and prediction' in J.-A. Meyer and H. Roitblat (eds): *Comparative Approaches to Cognition*. Cambridge MA: MIT Press, pp. 271–302.

Cheng, P. W. and **L. R. Novick.** 1990. 'A probabilistic contrast model of causal induction,' *Journal of Personality and Social Psychology* 58: 545–67.

Christiansen, M. H. and **N. Chater** (eds). 2001. *Connectionist Psycholinguistics*. Westport, CO: Ablex.

Corder, S. P. 1967. 'The significance of learners' errors,' *International Review of Applied Linguistics* 5: 161–9.

DeKeyser, R. 2001. 'Automaticity and automatization' in P. Robinson (ed.): *Cognition and Second Language Acquisition*. Cambridge: Cambridge University Press.

Ebbinghaus, H. 1885. *Memory: A Contribution to Experimental Psychology* (H.A.R.C.E.B. (1913), Trans.). New York: Teachers College, Columbia.

Ellis, N. C. 1998. 'Emergentism, connectionism and language learning,' *Language Learning* 48/4: 631–64.

Ellis, N. C. 2002a. 'Frequency effects in language processing: A review with implications for theories of implicit and explicit language acquisition,' *Studies in Second Language Acquisition* 24/2: 143–88.

Ellis, N. C. 2002b. 'Reflections on frequency effects in language processing,' *Studies in Second Language Acquisition* 24/2: 297–339.

Ellis, N. C. 2003. 'Constructions, chunking, and connectionism: The emergence of second language structure' in C. Doughty and M. H. Long (eds): *Handbook of Second Language Acquisition*. Oxford: Blackwell.

Ellis, N. C. 2005. 'At the interface: Dynamic interactions of explicit and implicit language knowledge,' *Studies in Second Language Acquisition* 27: 305–52.

Ellis, N. C. (2006). 'Selective attention and transfer phenomena in L2 acquisition: Contingency, cue competition, salience, interference, overshadowing, blocking, and perceptual learning,' *Applied Linguistics* 27: 2.

Ellis, N. C. (in press). 'Meta-analysis, human cognition and language learning' in J. Norris

and L. Ortega (eds): *Synthesizing Research on Language Learning and Teaching*. Amsterdam: John Benjamins.

Ellis, N. C. and **R. Schmidt,** 1998. 'Rules or associations in the acquisition of morphology? The frequency by regularity interaction in human and PDP learning of morphosyntax,' *Language & Cognitive Processes* 13/2&3: 307–36.

Elman, J. L., E. A. Bates, M. H. Johnson, A. Karmiloff-Smith, D. Parisi, and **K. Plunkett.** 1996. *Rethinking Innateness: A Connectionist Perspective on Development*. Cambridge, MA: MIT Press.

Farley, J. 1987. 'Contingency learning and causal detection in Hermissenda: 1. Behavior,' *Behavioral Neuroscience* 101: 13–27.

Gallistel, C. R. 2003. 'Conditioning from an information processing perspective,' *Behavioural Processes* 61: 1–13.

Gigerenzer, G. and **U. Hoffrage.** 1995. 'How to improve Bayesian reasoning without instruction: Frequency formats,' *Psychological Review* 102: 684–704.

Hasher, L. and **W. Chromiak.** 1977. 'The processing of frequency information: An automatic mechanism?' *Journal of Verbal Learning and Verbal Behavior* 16: 173–84.

Hodgson, J. M. 1991. 'Informational constraints on pre-lexical priming,' *Language and Cognitive Processes* 6: 169–205.

James, W. 1890. *The Principles of Psychology* Vol. 1. New York: Holt.

Jaynes, E. T. 1996. 'Probability theory with applications in science and engineering,' from http://bayes.wustl.edu/etj/science.pdf.html.

Jiménez, D. A. and **C. Lin.** 2002. 'Neural methods for dynamic branch prediction,' *ACM Transactions on Computer Systems* 20: 369–97.

Jurafsky, D. 2002. 'Probabilistic modeling in psycholinguistics: linguistic comprehension and production' in R. Bod, J. Hay, and S. Jannedy (eds): *Probabilistic Linguistics*. Harvard, MA: MIT Press, pp. 39–96.

Jurafsky, D. and **J. H. Martin.** 2000. *Speech and Language Processing: An Introduction to Natural Language Processing, Speech Recognition, and Computational Linguistics*. Englewood Cliffs, NJ: Prentice-Hall.

Kahneman, D. and **A. Tversky.** 1972. 'Subjective probability: A judgment of representativeness,' *Cognitive Psychology* 3: 430–54.

Kamin, L. J. 1968. ' ''Attention-like'' processes in classical conditioning' in M. R. Jones (ed.): *Miami Symposium on the Prediction of Behavior: Aversive Stimulation*. Miami, FL: University of Miami Press, pp. 9–31.

Kamin, L. J. 1969. 'Predictability, surprise, attention, and conditioning' in B. A. Campbell and R. M. Church (eds): *Punishment and Aversive Behavior*. New York: Appleton-Century-Crofts, pp. 276–96.

Kelley, H. H. 1967. 'Attribution theory in social psychology' in D. Levin (ed.): *Nebraska Symposium of Motivation* Vol. 15. Lincoln: University of Nebraska Press.

Kirsner, K. 1994. 'Implicit processes in second language learning' in N. C. Ellis (ed.): *Implicit and Explicit Learning of Languages*. San Diego, CA: Academic Press, pp. 283–312.

Klein, W. 1998. 'The contribution of second language acquisition research,' *Language Learning* 48: 527–50.

Krames, L., J. Graham, and **T. Alloway.** 1997. Sniffy: The virtual rat 4.5 for Windows. Portland, OR: Brooks/Cole.

Langacker, R. W. 1987. *Foundations of Cognitive Grammar: Vol. 1. Theoretical Prerequisites*. Stanford, CA: Stanford University Press.

Langacker, R. W. 2000. 'A dynamic usage-based model' in M. Barlow and S. Kemmer (eds): *Usage-based Models of Language*. Stanford, CA: CSLI Publications, pp. 1–63.

Lee, J. K. F. and **A. J. Smith.** 1984. 'Branch prediction strategies and branch target buffer design,' *IEEE Computer* 17: 6–22.

MacKay, D. 2004. The Dasher project, from http://www.inference.phy.cam.ac.uk/dasher/.

Mackintosh, N. J. 1975. 'A theory of attention: Variations in the associability of stimuli with reinforcement,' *Psychological Review* 82: 276–98.

MacWhinney, B. 1987. 'The competition model' in B. MacWhinney (ed.): *Mechanisms of Language Acquisition*. pp. 249–308.

MacWhinney, B. 1997. 'Second language acquisition and the Competition Model' in A. M. B. De Groot and J. F. Kroll (eds): *Tutorials in Bilingualism: Psycholinguistic Perspectives*. pp. 113–42.

MacWhinney, B. 2001a. 'The competition model: The input, the context, and the brain' in P. Robinson (ed.): *Cognition and Second Language Instruction*. New York: Cambridge University Press, pp. 69–90.

MacWhinney, B. 2001b. 'Emergentist approaches to language' in J. Bybee and P. Hopper (eds): *Frequency and the Emergence of Linguistic*

Structure. Amsterdam, Netherlands: Benjamins, pp. 449–70.

MacWhinney, B. and **E. Bates.** 1989. *The Cross-linguistic Study of Sentence Processing*. Cambridge: Cambridge University Press.

MacWhinney, B., E. Bates, and **R. Kliegl.** 1984. 'Cue validity and sentence interpretation in English, German, and Italian,' *Journal of Verbal Learning & Verbal Behavior* 23/2: 127–50.

MacWhinney, B., C. Pleh, and **E. Bates.** 1985. 'The development of sentence interpretation in Hungarian,' *Cognitive Psychology* 17/2: 178–209.

McDonald, J. L. 1986. 'The development of sentence comprehension strategies in English and Dutch,' *Journal of Experimental Child Psychology* 41: 317–35.

McDonald, J. L. and **B. MacWhinney.** 1991. 'Levels of learning: A comparison of concept formation and language acquisition,' *Journal of Memory & Language* 30/4: 407–30.

Manning, C. D. 2003. 'Probabilistic syntax' in R. Bod, J. Hay, and S. Jannedy (eds): *Probabilistic Linguistics*. Cambridge, MA: MIT Press, pp. 289–341.

Manning, C. D. and **H. Schuetze.** 1999. *Foundations of Statistical Natural Language Processing*. Cambridge, MA: The MIT Press.

Marslen-Wilson, W. 1990. 'Activation, competition, and frequency in lexical access' in G. T. M. Altmann (ed.): *Cognitive Models of Speech Processing*. Cambridge, MA: ACL-MIT Press, pp. 148–172.

Matessa, M. and **J. R. Anderson.** 2000. 'Modeling focused learning in role assignment,' *Language & Cognitive Processes* 15/3: 263–92.

Newell, A. 1990. *Unified Theories of Cognition*. Cambridge, MA: Harvard University Press.

Newell, A. and **P. Rosenbloom.** 1981. 'Mechanisms of skill acquisition and the law of practice' in J. Anderson (ed.): *Cognitive Skills and Their Acquisition*. Hillsdale, NJ: Lawrence Erlbaum Associates, pp. 1–55.

Norman, D. 1988. *The Psychology of Everyday Things*. New York: Basic Books (Perseus).

Norman, D. 1993. *Things That Make Us Smart: Defending Human Attributes in the Age of the Machine*. Cambridge, MA: Perseus Publishing.

Oldfield, R. and **A. Wingfield.** 1965. 'Response latencies in naming objects,' *Quarterly Journal of Experimental Psychology A* 17/4: 273–81.

Pearce, J. M. and **M. E. Bouton.** 2000. 'Theories of associative learning in animals,' *Annual Review of Psychology* 52: 111–39.

Perdue, C. (ed.). 1993. *Adult Language Acquisition: Crosslinguistic perspectives*. Cambridge: Cambridge University Press.

Peterson, C. R. and **L. R. Beach.** 1967. 'Man as an intuitive statistician,' *Psychological Bulletin* 68: 29–46.

Plaut, D. C., J. L. McClelland, M. S. Seidenberg, and **K. Patterson.** 1996. 'Understanding normal and impaired word reading: Computational principles in quasi-regular domains,' *Psychological Review* 94: 523–68.

Rescorla, R. A. 1968. 'Probability of shock in the presence and absence of CS in fear conditioning,' *Journal of Comparative and Physiological Psychology* 66: 1–5.

Rescorla, R. A. and **A. R. Wagner.** 1972. 'A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement' in A. H. Black and W. F. Prokasy (eds): *Classical Conditioning II: Current Theory and Research*. New York: Appleton-Century-Crofts, pp. 64–99.

Reynolds, G. S. 1961. 'Attention in the pigeon,' *Journal of the Experimental Analysis of Behavior* 4: 203–8.

Robinson, P. (ed.). 2001. *Cognition and Second Language Instruction*. Cambridge: Cambridge University Press.

Rumelhart, D. E. and **J. L. McClelland.** (eds). 1986. *Parallel Distributed Processing: Explorations in the Microstructure of Cognition* Vol. 2: Psychological and biological models. Cambridge, MA: MIT Press.

Sampson, G. 2001. *Empirical Linguistics*. London: Continuum.

Schmidt, R. 1984. 'The strengths and limitations of acquisition: A case study of an untutored language learner,' *Language, Learning, and Communication* 3: 1–16.

Schmidt, R. 2001. 'Attention' in P. Robinson (ed.): *Cognition and Second Language Instruction*. Cambridge: Cambridge University Press, pp. 3–32.

Schmitt, N. (ed.). 2004. *Formulaic Sequences: Acquisition, Processing and Use*. Amsterdam: John Benjamins.

Schooler, L. J. 1993. *Memory and the Statistical Structure of the Environment*. Carnegie Mellon University, Pittsburgh, PA.

Schooler, L. J. and **Anderson, J. R.** 1997. 'The role of process in the rational analysis of memory,' *Cognitive Psychology* 32/3: 219–50.

Schumann, J. H. 1978. *The Pidginisation Process: A Model for Second Language Acquisition*. Rowley, MA: Newbury House.

Sedlmeier, P. and **T. Betsc.** 2002. *Etc.—Frequency Processing and Cognition*. Oxford: Oxford University Press.

Sedlmeier, P. and **G. Gigerenzer.** 2001. 'Teaching Bayesian reasoning in less than two hours,' *Journal of Experimental Psychology: General* 130: 380–400.

Shanks, D. R. 1995. *The Psychology of Associative Learning*. New York: Cambridge University Press.

Shannon, C. E. 1948. 'A mathematical theory of communication,' *Bell Systems Technological Journal* 27: 623–56.

Sinclair, J. 1991. *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.

Slobin, D. I. 1996. 'From ''thought and language'' to ''thinking for speaking'' ' in J. J. Gumperz and S. C. Levinson (eds): *Rethinking Linguistic Relativity*. Cambridge: Cambridge University Press.

Tomasello, M. (ed.). 1998. *The New Psychology of Language: Cognitive and Functional Approaches to Language Structure*. Mahwah, NJ: Erlbaum.

Tomasello, M. 2003. *Constructing a Language*. Boston, MA: Harvard University Press.

Tversky, A. and **D. Kahneman.** 1982. 'Evidential impact of base rates' in D. Kahneman, P. Slovic, and A. Tversky (eds): *Judgment under Uncertainty: Heuristics and Biases*. Cambridge: Cambridge University Press, pp. 153–160.

Wagner, A. R., F. A. Logan, K. Haberlandt, and **T. Price.** 1968. 'Stimulus selection in animal discrimination learning,' *Journal of Experimental Psychology* 76: 171–80.

Ward, D. J. and **D. J. C. MacKay.** 2002. 'Fast hands-free writing by gaze direction,' *Nature* 418: 838.

Wasserman, E. A., S. M. Elek, D. L. Chatlosh, and **A. G. Baker.** 1993. 'Rating causal relations: The role of probability in judgments of response-outcome contingency,' *Journal of Experimental Psychology: Learning, Memory, & Cognition* 19: 174–98.

Widrow, B. and **M. E. Hoff.** 1960. 'Adaptive switching circuits,' *1960 IRE WESCON Convention Record* (Pt. 4): 96–104.

Williams, J. N. 1996. 'Is automatic priming semantic?' *European Journal of Cognitive Psychology* 22: 139–51.

Wixted, J. T. and **E. Ebbesen.** 1991. 'On the form of forgetting,' *Psychological Science* 2: 409–15.

Yudkowsky, E. 2003. 'An Intuitive Explanation of Bayesian Reasoning' from http://yudkowsky.net/bayes/bayes.html.