Running head: SEMANTIC CONTEXT EFFECTS

A distributional model of semantic context effects in lexical processing

Scott A. McDonald

University of Edinburgh, Edinburgh, Scotland

Chris Brew

The Ohio State University, Columbus, Ohio.

Address for correspondence:

Scott McDonald
Department of Psychology
University of Edinburgh
7 George Square
Edinburgh EH8 9JZ
Scotland, UK

Tel: +44 131 650 3423
Fax: +44 131 650 3461
Email: Scott.McDonald@ed.ac.uk

15,900 words

Abstract

One of the most robust findings of experimental psycholinguistics is that the context in which a word is presented influences the effort involved in processing that word. We present a novel model of contextual facilitation based on word co-occurrence probability distributions, and empirically validate the model through simulation of three representative types of context manipulation: single word priming, multiple-priming and contextual constraint. In our simulations the effects of semantic context are modeled using general-purpose techniques and representations from multivariate statistics, augmented with simple assumptions reflecting the inherently incremental nature of speech understanding. The contribution of our study is to show that special-purpose mechanisms are not necessary in order to capture the general pattern of the experimental results, and that a range of semantic context effects can be subsumed under the same principled account.

A distributional model of semantic context effects in lexical processing

Introduction

Evidence from a range of experimental paradigms has shown that the context in which a word is presented influences the effort involved in processing that word. Words are recognized more quickly and more accurately if the immediately preceding context is related in meaning, compared with an unrelated or a "neutral" context. This preceding context (or prime) can consist of a single word, a set of words, or a sentence fragment that is, in some approximate sense, related to the target word. We will refer to such contexts as semantically compatible with the target. The standard experimental procedure is to hold the target word constant while manipulating the context.

The facilitatory effect of a semantically compatible context on lexical processing effort is one of the most robust findings in psycholinguistics, and many different techniques are available to probe its properties, including visual and auditory lexical decision tasks (e.g., Meyer & Schvaneveldt, 1971; Moss, Ostrin, Tyler & Marslen-Wilson, 1995), pronunciation (e.g., Lupker, 1984), event-related brain potentials (Brown, Hagoort & Chwilla, 2000), and the recording of eye movements during normal reading (e.g., Duffy, Henderson & Morris, 1989). The extensive literature concerned with contextual influences on lexical processing divides into three main strands: (1) lexical priming (single-word contexts, where the prime-target relation is semantic or associative in nature); (2) multiple priming (two or more individual lexical primes); and (3) contextual constraint (the set of primes is structured by linguistic relationships with one another). An example of this last strand is the use of a sentence frame that is compatible/incompatible with its final word (e.g., "On a hot summer's day many people go to the beach"/"He scraped the cold food from his beach").

The assumption (often left implicit) is that contextual priming, in its various forms, can be ascribed to fundamental mechanisms of cognition. Because these effects are robust and apparently automatic, researchers often seek explanations in terms of low-level mechanisms such as spreading activation (Anderson, 1983; Collins & Loftus,

1975), compound-cue models (Ratcliff & McKoon, 1988), and distributed neural network models (Cree, McRae & McNorgan, 1999; Plaut, 1995). When these relatively simple models fail to cover every aspect of the behavioral data, one response has been to develop theories that meld several mechanisms (Keefe & Neely, 1990). Another response is to prefer simplicity over detailed explanatory power. Plaut and Booth (2000), for example, make no claim about their network model's ability to account for blocking and strategy effects, arguing that it would detract from the main point of their work to focus on these, which may in any case be due to other mechanisms.

We present a model even simpler than Plaut and Booth's. This choice is inspired by the rational analysis approach to understanding cognition (for an overview, see Chater & Oaksford, 1999; Oaksford & Chater, 1998). We demonstrate that distributional information available from the linguistic environment – information about word usage that is inherent in large language corpora – can capture salient aspects of a range of data from all three strands of the literature. The most important aspect of our analysis is the breadth of coverage which can achieved by a single simple explanatory idea. It is not necessary to invoke distinct mechanisms for the different priming settings, nor are there independently tunable parameters for different purposes.

The core of our work is a computational model of the process of constructing semantic expectations about upcoming words – a model that uses simple mechanisms and standard statistical tools to tap into the distributional information present in the linguistic environment.

Expectations about Meaning

A number of human cognitive behaviors – such as memory retrieval and categorization – have been claimed to be adapted to their environment (see, e.g., Anderson, 1990). The normal setting for speech processing is an environment in which acoustic cues are unreliable or absent, so it makes sense for the hearer to draw upon available resources in order to maximize the chances of successful comprehension. Such

resources include any prior knowledge that the hearer might have about what the speaker will say next.

One way to encode prior knowledge is to construct probabilistically weighted hypotheses about the meaning of upcoming words. Our model is of this type. Specifically, our system maintains a vector of probabilities as its representation of the current best guess about the likely location in semantic space of the upcoming word. When that word is observed, the system updates its meaning representation to reflect the newly arrived information. The update mechanism, which uses standard multivariate distributions from Bayesian statistics, is designed to give greater weight to recent words than to those far in the past. Such a discounting mechanism has no independent justification in terms of probability theory, but is obviously necessary if the model is to capture the dynamic behavior of utterances. Without it, all that would be learnt would be a very stable estimate of the central position of the semantic space defined by the entire working vocabulary of the language in question.

The hypothesis that language comprehension involves an automatic process of expectation-building is by no means new. For instance, Tanenhaus and Lucas (1987) propose that the representation of the current discourse developed by the listener, together with the local context and world knowledge "… can often be used to generate reasonable expectations about the incoming speech" (p. 224). They suggest that this strategy could increase lexical processing efficiency in context if the structure of the mental lexicon was similar to the structure of the environment. What is new in our model is a fully explicit implementation of this idea. Note that our use of the term "expectation" is only distantly related to the term "expectancy" found in the semantic priming literature. We are talking about probabilistically weighted hypotheses, not the expectancy sets discussed in previous work in semantic priming (Becker, 1980; Keefe & Neely, 1990; Neely, 1991). The claim there is that an expectancy set is generated using the prime word to explicitly create a set of potential targets; if this set does not contain the target word actually encountered, processing is inhibited. We argue for a

weaker claim: probabilistically weighted hypotheses can be used to predict the region of semantic space in which the target is likely to appear.

Distributional Approaches to Word Meaning

Semi-automatic methods for representing word meaning have been developed over the last few years; these data-intensive techniques use statistical information about the contexts in which a word occurs as a representation for the word itself (e.g., Lund & Burgess, 1996; Redington, Chater & Finch, 1998). In this approach, a word is represented in terms of its relationships with other words, where "relationship" is defined as simple co-occurrence – within a sentence or within a "context window" of a fixed size. For example, the frequency with which "context words" such as good and true co-occur with value contribute towards distinguishing the meaning of value from other words. This approach is made practical by the availability of large corpora of electronically accessible text.

As in more traditional feature-based theories of word meaning (where features are either predefined [e.g., Plaut & Shallice, 1994; Smith & Medin, 1981] or generated using a norming procedure [McRae, de Sa & Seidenberg, 1997]), all the relevant information about a word can be collected into a vector, and viewed as a point in high-dimensional "semantic space". But in our approach the frequency with which a word occurs with other words in the linguistic environment is the only information used to position it in the space. To the extent that a corpus mirrors the language input available to the typical hearer, semantic representations constructed from co-occurrence statistics over that corpus summarize regularities which the speaker might be able to exploit. They provide information about word usage and – if one subscribes to contextual theories of meaning (e.g., Wittgenstein, 1958) – meaning.

A number of studies have tried to uncover correlations between the similarity structure of word vectors and measurable indicators of human performance, such as lexical priming (e.g., Lund, Burgess & Atchley, 1995; McDonald & Lowe, 1998) and

semantic similarity ratings (McDonald, 2000). The same representations also play a role in simulations of children's vocabulary acquisition and synonym choice tests (Landauer & Dumais, 1997). All of these studies rely on the basic assumption that word vectors can function as convenient proxies for more highly articulated semantic representations.

Our primary claim is that word vectors also provide a compact and perspicuous account of priming phenomena previously ascribed to a multitude of mechanisms. We do not wish or need to claim any psychological reality for distribution-based mechanisms (although steps in this direction have been taken by Landauer and Dumais, 1997).

Contextual Facilitation

Contextual facilitation refers to a measurable reduction in processing effort resulting from presenting a target word in a semantically compatible context, compared with presenting the same word in an incompatible context. The size of the semantic context effect is thus related to the difference in processing difficulty (as measured by dependent variables such as response time and accuracy) that can be attributed to the manipulation of semantic context.

Semantic context, as exploited by humans, is probably very rich: it could extend to a full model of the current discourse. This would draw upon knowledge about the world, integrating this knowledge with the properties of the referents of the words in the context and features of the communicative environment. We are not in a position to model such knowledge, however desirable it might be. Our approach uses nothing more than the frequencies of nearby words. This is admittedly simplistic, but has two major advantages: (1) it allows fully explicit computational modeling; and (2) it holds in check the number of independent variables which would otherwise need to be manipulated.

In the following, we summarize the central features of the three types of context manipulation examined in the present paper, and consider how the our model of semantic context effects would account for the relevant empirical data.

Single-word Priming

The single word lexical priming paradigm relies on the manipulation of a single context word. Usually a relationship in meaning between the prime word (e.g., value) and the target word (e.g., worth) facilitates responses made to the target. Responses (e.g., lexical decisions, pronunciation) are quicker and more accurate when the prime is related to the target than when it is unrelated (Meyer & Schvaneveldt, 1971; for a review, see Neely, 1991). The paradigm is therefore used when the intent is to uncover the structure of the mental lexicon, which is assumed to encode semantic relationships.

Traditional spreading activation models (e.g., Collins & Loftus, 1975) explain automatic lexical priming in representational terms: encountering the prime word activates its conceptual (or semantic) entry in memory, and this activation is assumed to spread to other, semantically related representations. These spreading activation models make the prediction that a target word that maps to one of these "pre-activated" entries will be recognized faster than a word whose conceptual representation is not primed. Thus, the account postulates the existence of links between concepts, along with a mechanism that spreads activation from one concept to another. If conceptual representations are directly associated with individual words, as the spreading activation literature appears to assume, the model describes the direct prediction of one word on the basis of another. Spreading activation accounts of lexical priming also assume that links between concepts vary in strength, with variation corresponding to the relatedness or associative/predictive strength between nodes.

A predictive relationship between prime and target appears to be sufficient to produce priming (e.g., Lupker, 1984); if priming were a simple reflection of lexical prediction – predicting the identity of the target word given knowledge of the prime – then one would expect a correlation between effect size and predictability. Later studies (Hodgson, 1991) showed that predictability (as measured by normed association strength) was uncorrelated with the magnitude of the priming effect. Under our model the prime does not directly activate concepts, but instead gives rise to an expectation

about the location in semantic space of upcoming words. When a word, whether expected or unexpected, actually arrives, the model generates a corresponding representation of its true location in semantic space, and compares the expected and the true locations. The hearer can be though of as integrating predictive and perceptual evidence to track a trajectory as it develops in high-dimensional semantic space. This parallels the use that an athlete makes of visual feedback in updating best-guess projections of the likely trajectory of a moving ball, although the timescale of the process is very different. If one accepts the parallel, our account considers single-word priming as a facet of a generally useful predictive mechanism. We regard this parsimony as an advantage (although see Sharkey and Mitchell [1985] for an opposing view).

Multiple Priming

In the multiple priming paradigm (also known as "summation priming"), the prime consists of two or more words related to the target. Several researchers (e.g., Brodeur & Lupker, 1994; Balota & Paul, 1996) have demonstrated a multiple-prime advantage: two or more related prime words triggered a larger overall priming effect than a single related prime. For example, Balota and Paul (1996) found that the prime sequence <copper, bronze> primed metal to a significantly greater degree than a prime pair consisting of either copper or bronze coupled with an unrelated word such as drama. Note that while the prime words are presented in sequence, they do not themselves form a syntactically coherent unit.

Contextual Constraint

Contextual constraint refers to the influence of the local linguistic context on the predictability of a word presented in that context. In other words, the more strongly the target word is constrained by the context, the greater the likelihood that the target will be observed. Contextual constraint has been manipulated as an independent variable in a large number of studies using experimental paradigms such as lexical decision (e.g.,

Schwanenflugel & Shoben, 1985), naming (Nebes, Boller & Holland, 1986), rapid serial visual presentation (e.g., Altarriba, Kroll, Sholl & Rayner, 1996), event-related brain potentials (e.g., Van Petten, Coulson, Rubin, Plante & Parks, 1999), and the recording of eye movements during reading (e.g., Rayner & Well, 1996). The standard finding is that words are processed more rapidly when preceded by a constraining context than by an unconstraining (or neutral) context (e.g., Altarriba et al., 1996; Schwanenflugel & Shoben, 1985).

As in multiple priming, a sequence of prime words are presented before the target, but this time the primes are composed into a coherent syntactic context. This could affect lexical processing in at least two ways. One possibility is that the processor will be able to exploit the newly available syntactic structure, using reasoning about the relations between the prime words to make inferences about the target word. The second possibility is that predictions about the target are generated from the context words independently of any syntactic dependencies (evidence compatible with this is provided by Masson, 1986; but see Morris, 1994).

The explanation for contextual constraint effects has usually been framed in terms of a feature-based theory of meaning. For example, according to Schwanenflugel and LaCount's (1988) Feature Restriction Model a constraining sentence context is a context that imposes detailed semantic feature restrictions on upcoming words. Processing ease is taken to mirror the ability of the upcoming word to meet these feature restrictions. Under the Feature Restriction Model semantic compatibility reduces to a constraint-satisfaction process; facilitation occurs when the semantic features associated with the upcoming word conform to the feature restrictions imposed by the context. Notice that a simulation of such a theory would require us to specify not only the feature repertoire but also the details of what it means for a word to be compatible with a context. There is therefore scope for considerable variation within implementations of this type of model. Because the model we present uses distributional information directly, it does not need a pre-specified feature set, and because it replaces the test for feature compatibility with the weaker notion of proximity in semantic space, we are freed

from the need to invent ad hoc mechanisms for quantifying compatibility. It remains the case that our model could at a pinch be regarded as an instance of the feature restriction model, albeit one with a very particular choice of feature repertoire and compatibility mechanism.

Notice that neither in our model or Schwanenflugel and LaCount's is there an explicit role for syntax. As far as the mechanism is concerned, a syntactically coherent context is simply a particular set of prime words presented in a particular order. Our model is order-dependent, in the sense that it yields different results when prime words are presented in alternative orders, but it is not syntax-dependent. The model will only be sensitive to syntactic manipulations if they affect the identity and/or order of the prime words that are presented to the mechanism. This is a limitation of our model and an avenue for further research.

Incremental Interpretation

Spoken language comprehension is constrained by the timeline – utterances are produced (or at least articulated) serially, and it seems clear that understanding also proceeds incrementally, or word-by-word (e.g., Marslen-Wilson & Tyler, 1980). There is an appealing functional motivation for this incrementality: if partial meanings are available before the end of the clause or sentence, the processor can more easily cope with "noisy" input (due to auditory interference, disfluency, etc.). To the extent that the speaker's communicative intent can be recovered, at least in part, without reliance on a completed syntactic analysis, transmission of information from speaker to hearer will be more efficient. Under our working assumptions, predictions about the meaning of upcoming words could be created and updated as each word of the input is processed. As words become available from the unfolding input, they are integrated into the processor's expectations about meaning, with the effect that predictions that started out vague become increasingly precise.

There is compelling experimental evidence supporting the incrementality of semantic processing (e.g., Altmann & Kamide, 1999; Muckel, Scheepers, Crocker &

Müller, 2002;Van Berkum, Hagoort & Brown, 1999). Using an eye-tracking paradigm to investigate the processing of simple spoken sentences such as "the boy will eat the cake", Altmann and Kamide (1999) demonstrated that information carried by the verb eat could restrict the set of possible (visual) referents of its direct object, even before the direct object was realized in the auditory input. In an event-related brain potential (ERP) study, Van Berkum et al. (1999) found very rapid disruptive effects for words that were semantically anomalous in either the sentence or discourse context, suggesting that the meanings of individual words are integrated into a discourse-level representation as the input unfolds. Both findings are consistent with the view that comprehension involves a process of building expectations about meaning. In Altmann and Kamide's (1999) study the preceding context (i.e., the subject noun and verb) can be seen as supplying prior knowledge about the meaning of its direct object. Subjects were more likely to fixate on an item in their visual environment that was semantically compatible with the verb when the verb generated detailed expectations for its direct object (e.g., eat), compared with a verb that generates less detailed expectations (e.g., move).

It seems clear that language comprehension is an incremental process, so a parsimonious model of lexical processing should accommodate the incremental nature of semantic interpretation. The model presented below satisfies these concerns.

The ICE Model

In this section, we describe our implemented computational model of contextual facilitation. Before launching into a detailed description of the ICE model (for Incremental Construction of semantic Expectations), we summarize its fundamental functionality. The ICE model simulates the difference in processing cost between circumstances where a target word is preceded by a related prime and where the same target is preceded by an unrelated prime, by quantifying the ability of the distributional characteristics of the word(s) in the prime to predict the distributional properties of the target. This prediction – a probabilistic representation of the expected meaning of the

target word – is constructed by combining the prior knowledge provided by the distributional information associated with each prime word in an incremental fashion.

The Influence of Prior Knowledge

Previous context, even a single word, is often informative about the identity, and consequently the meaning of a given word w. For example, the adverb hermetically is a very strong cue that the next word will be sealed. Even if we have no idea about the intrinsic meaning of hermetically,[1] we can predict that the next word will concern a state of closure. Intuitively, the prior knowledge that context has provided to the lexical processor should influence the effort involved in recovering the meaning of w – encountering hermetically should facilitate sealed. We can describe this anticipated effect of prior information either as an increase in processing efficiency, or as a reduction in the cost of lexical processing. For all practical purposes these descriptions are equivalent. The task of the ICE model is to estimate the cost of processing a given word w, taking into consideration available prior knowledge about its meaning.

Estimating Processing Cost

We make two working assumptions: (1) a significant portion of the variability in lexical processing behavior reflects the relative effort of recovering the meaning of a word from its phonological or orthographic form; and (2) the processor attempts to reduce the average cost of recovering word meaning by forming semantic expectations. Given these assumptions, the cost of recovering word meaning can be seen as a consequence of the "fit" between the expected meaning of a word and its "actual" meaning. In previous work (McDonald, 2000; McDonald & Shillcock, 2001a,b), we showed how this quantity can be estimated as the amount of information conveyed by a word about its contexts of use. This estimate of processing effort is derived entirely

[1] It is not clear to us that hermetically even needs to have an intrinsic meaning, since it only occurs in one highly specific context.

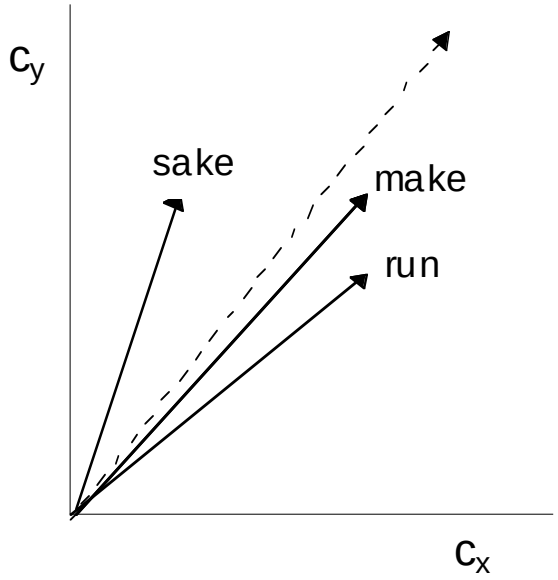

Figure 1. Example two-dimensional vector representations for the words sake, make and run (solid lines) and the null context (dashed line).

from lexical co-occurrence probabilities; it is computed from the distributional information inherent in large language corpora using the standard information-theoretic measure of relative entropy (our choice of this measure will become clear later).

Consider words that can be deployed in a wide variety of linguistic situations, like make and run. Such words reveal little about the situations and circumstances in which they appear. In contrast, words such as sake are extremely informative about their contexts of use; they occur only in fixed expressions such as for heaven's sake. Other words fall on the continuum between these extreme points. In previous work we have tested the utility of distributional data in quantifying this contextual property. Our assumption was that the more distinctive the distributional representation of a word (the it differs from the distributional representation of the "null context") the more difficult it is to process (McDonald & Shillcock, 2001b). We expect that the contextual distribution associated with sake will be sharply different from the null context, whereas the distributions associated with make and run will be much less distinguished. The underlying idea can be explained visually, provided that one temporarily accepts the

restriction of working in a semantic space where there are only two context words, c_x and c_y, and thus two dimensions.

In Figure 1, target words are shown as solid vectors starting at the origin. The end points are determined by the number of times that each context word has occurred with each target word. Counts for the first context word (c_x) determine the x-coordinates of end points while counts for the second (c_y) determine the y-coordinates. The diagram shows a semantic space in which make and run happen to have both occurred the same number of times with c_x, and that sake and make both appear the same number of times with c_y. One way to measure the similarity of the vectors representing two words is to focus on their directions: the direction of the vector for make is determined by the ratio between the number of times that make has occurred with each of the two context words. The dotted arrow represents the direction selected by the ratios of the occurrence frequencies of c_x and c_y in the whole corpus. If one repeatedly and randomly dives into the corpus, recording the number of times that one sees c_x and c_y near to the current location, the indicated direction is the best guess for the ratio between the number of times that one will observe each of the context words. This is the distributional representation of the null context. To the extent that distributional representations are useful as proxy representations of meaning, the dotted line can stand for the (doubtless very vague and unspecific) "meaning" of the null context, while the distributional representations corresponding to the solid lines can stand for the meanings of each of the three target words.

Under these assumptions, we can see that make and run are closer to each other than is sake to either, and that both make and run are close to the null context. Therefore we predict that it will be harder to process sake in the null context than it is to process the other words.

The argument in the previous paragraphs appealed to visual intuition as a measure of similarity between distributions. A more precise measure is needed. Much work in information retrieval has used the cosine of the angle between the vector representations of two objects as a measure of similarity, but we wanted a probabilistic measure, because

these are easier to extend and reason about, so chose to use relative entropy (also known as Kullback-Leibler divergence).

The relative entropy D(**p**||**q**) of an estimated discrete probability distribution **q** with respect to a true distribution **p** is defined as:

$$D(p\|q) = \sum p \cdot \log \frac{p}{q} \tag{1}$$

This quantity is always non-negative, zero only when **p**=**q**, and quantifies the difference between two discrete probability distributions over the same domain. It gives a precise information-theoretic interpretation of the informal notions of distance and direction used above.

In McDonald and Shillcock (2001b), we used relative entropy to compare the distributional representation of the null context with the distributional representation of the target word. Lexical decision response times correlated with the relative entropy of the distributional representation of the target word measured with respect to the distributional representation of the null context. The present paper extends this to the multi-word setting. We retain the fundamental idea that processing effort is related to the difference between successive estimates of a location in semantic space. Intuitively, less information about the meaning of sealed should be conveyed when it is encountered immediately after hermetically, compared with the case where sealed is preceded by a neutral, or less informative cue such as carefully.

Until now it has been sufficient to represent regions in semantic space by unit direction vectors, which pick out the central point of the region. But in order to progress, we need to augment this measure of central tendency with a measure of dispersion. The intuition is that the processor is sometimes more and sometimes less committed to a hypothesis, so we want a means of representing these degrees of commitment. In other words, we need a way of expressing probability distributions over vector representations. The next section provides formal apparatus to do this.

Formalizing the model

We now formalize the intuitions developed above. The goal is to provide an account of the way in which words and their contexts interact. For this purpose, we idealize the process of context formation as the outcome of a series of independent draws from a multinomial distribution. The multinomial distribution is fully specified by a set of parameters $\underline{\theta} = (\underline{\theta}_1 \ldots \underline{\theta}_k)$, which sum to 1. Given any setting of $\underline{\theta}$, we can calculate the likelihood that a particular set of context words $c_i \ldots c_k$ will be observed.

$$P(\theta|c) = \frac{P(c|\theta)P(\theta)}{P(c)} = \frac{P(c|\theta)P(\theta)}{\int P(c|\theta)P(\theta)d\theta} \quad (2)$$

Although this expression is exact and fully general, the posterior density is, in general, difficult to evaluate, because of the integral over a high-dimensional space in the denominator. Historically, this has been a major obstacle to the adoption of the Bayesian approach in applications which (unlike our own) require the full posterior distribution. Fortunately, a convenient solution is available if one accepts a prior distribution that is conjugate to the likelihood function. In this case, by the definition of conjugacy, the posterior distribution will take on the same functional form as the prior. To use this idea we have to choose appropriate prior and posterior distributions. We now introduce and motivate the choices that we made for the ICE model:

$$P(c_1,...,c_k|\theta) = \frac{n!}{c_1!\ldots c_k!}\prod_{i=1}^{k}\theta_i^{c_i} \quad (3)$$

We could use multinomials as our meaning representations, but we want to represent more than just the maximum of the likelihood. For reasons of simplicity prefer distributions that have convenient analytical properties and concise parametric

representations[2]. One such is the Dirichlet distribution, which is widely used in Bayesian statistics (Gelman, Carlin, Stern & Rubin, 1995). The Dirichlet distribution is conjugate to the familiar multinomial. This means that if we begin with prior information expressed in the form of a Dirichlet, then update it with data drawn from a multinomial, the posterior distribution can also be expressed as a Dirichlet, albeit one whose parameters have been adjusted to better fit the recently observed data. This closure property is crucial to our application, since it allows us to model both prior and posterior hypotheses in the same way. The Dirichlet distribution has the form:

$$P(\theta) = \frac{1}{B(\alpha_1 \ldots \alpha_k)} \prod_{i=1}^{k} \theta_i^{\alpha_i - 1} \tag{4}$$

Because it is a conjugate prior for the multinomial distribution, the Dirichlet can easily be combined with the multinomial expression for the likelihood (see Gelman, Carlin, Stern & Rubin, 1995, p.76) to specify the posterior distribution of $\underline{\theta}$:

$$P(\theta | c_1, \ldots, c_k) = \frac{1}{B(c_1 + \alpha_1, \ldots, c_k + \alpha_k)} \cdot \frac{n!}{c_1! \ldots c_k!} \cdot \prod_{i=1}^{k} \theta_i^{c_i + \alpha_i - 1} \tag{5}$$

The crucial difference between the Dirichlet and the multinomial is that the latter is parameterized by a vector of probabilities, while the former is specified by a vector of arbitrary real-valued weights. Because the probabilities must normalize, a multinomial is equivalent to a direction vector, while a Dirichlet has an extra degree of freedom, because the weights need not normalize. In our application (though not in general) the Dirichlet parameters are non-negative. We can therefore represent a Dirichlet as a combination of a direction vector and single positive scale parameter. The prior weights can be interpreted as representing prior knowledge about $\underline{\theta}$ in the form of "virtual samples". If

[2] For us this is a methodological choice, but until recently limitations on computing power have imposed severe restrictions on the distributions which could be used in large scale statistical modelling.,

$\sum_j \alpha_j$ is small, the posterior distribution will be highly responsive to new data, while if it is large the posterior will tend to stay close to the prior. Now consider the update process. The inputs to the process are the parameters α_i of the prior distribution and the components c_i of the vector representation of the incoming word:

$$\hat{\theta}_i = \frac{c_i + \alpha_i}{\sum_{j=1}^{k} (c_j + \alpha_j)} \tag{6}$$

Here, the values of c_i are simply the component values of the vector representation for c. The main motivation for using this Bayesian update procedure is the straightforward way in which prior knowledge can be combined with new data; with every iteration the posterior distribution serves as the prior distribution for the next cycle. By applying the update iteratively to successive words in an utterance, the prior distribution is revised on a word-by-word, incremental basis,

The final step is to compute the amount of information conveyed by w when preceded by context cue c. This involves calculating the relative entropy between the updated prior distribution $P(\theta)$ and the empirical distribution $P(\theta|w)$ (as encoded in the co-occurrence vector representation for w), resulting in a value expressed in units of information, or bits (equation (7) is identical to equation (1)):

$$D\big(P(\theta) \,||\, P(\theta|w)\big) = \sum_{i=1}^{n} P(\theta_i|w) \log_2 \frac{P(\theta_i|w)}{P(\theta_i)} \tag{7}$$

It is worth clarifying the model function by working through the hermetically sealed example. The first step of the procedure is to extract the co-occurrence information associated with both hermetically and sealed from the corpus, creating k-component vector representations for each word. Imagine that we have constructed a three-dimensional semantic space (i.e., k=3), and that the corpus frequencies of the three context words c_1, c_2 and c_3 are 150, 120 and 50, respectively. Imagine also that we have extracted the vector representations (14, 1, 10) for hermetically and (55, 4, 41) for sealed

(i.e., c_1 co-occurred 14 times with hermetically and 55 times with sealed). So, given the presence of hermetically, our task is to compute the difference between the expected meaning of the next word and its "actual" meaning (sealed), using the relative entropy measure.

Recall that $P(\theta)$ represents prior knowledge about the meaning of upcoming words. Because the shape of this distribution is initially unknown, we define $P(\theta)$ to be a Dirichlet distribution, and set the values of the Dirichlet parameters α_1, α_2, and α_3 to the corpus frequencies of the three context words (150, 120 and 50).

We now update this initial prior with the co-occurrence data for hermetically, giving the Dirichlet posterior $P(\theta|\text{hermetically})$. This updating simply involves revising each prior parameter θ_i using the equation for the posterior mean (5), resulting in the distribution (0.48, 0.35, 0.17). The last step is to calculate the ICE value for sealed as the relative entropy between the revised prior distribution (0.48, 0.35, 0.17) and the empirical distribution for sealed (0.55, 0.04, 0.41). The ICE value for this contrived example is 0.498; i.e., sealed carries 0.498 bits of information about its contexts of use when preceded by hermetically.

The above procedure presents a novel view of incremental semantic processing: if the prior distribution is iteratively revised to take into account the new evidence provided by each successive word[3] in an utterance, then estimates of processing effort become available on a word-by-word basis.

Overview of Simulations

The ICE model relates contextual facilitation to the amount of information a word conveys about its linguistic contexts of use, predicting that it is more costly for the

[3] More precisely, each successive content word: we assume that function words such as of and could do not contribute to the formation of semantic expectations (although they are predictive of lexical properties such as syntactic category), and therefore will not influence the form of the posterior distribution.

processor to incorporate a lot of information than a little. We compare the predictions of the ICE model against human behavioral data. A series of three computational experiments was conducted in order to assess the ability of the ICE model to account for a variety of published semantic context effects. These simulations involved applying the ICE model to the materials used in the original studies; thus they abstract over experimental differences such as the actual task employed, the dependent measure of lexical processing effort, and the relative timing of prime and target presentation. To the extent that the simulations reveal patterns consonant with those shown by human subjects we will have provided support for the claim that both the computational and the experimental studies are tapping into important properties of the task. Since these are re-analyses, we do not have access to all details of the original studies. In particular we do not have the response time data for individual items. It would be interesting to attempt a more detailed comparison, but that is left for further work.

Model Parameters

Corpus Distributional information was extracted from the spoken language component of the 100 million word British National Corpus (BNC). This subcorpus (henceforth BNC-spoken), containing 10.3 million words, is large enough to ensure that reliable co-occurrence statistics could be obtained for approximately 8,000 lexemes (McDonald & Shillcock, 2001a). We chose to use the domain of spoken language as a source of distributional information because spoken language constitutes a more realistic sample of the linguistic environment than written language. This is because, especially during early language development, people are exposed to much more speech than text .

We use a lemmatized version of the BNC-spoken. In this corpus, inflected forms of nouns (e.g., chairs), verbs (e.g., cried, cries, crying) and adjectives (e.g., stronger, strongest) were replaced with their canonical forms (chair; cry; strong). Lemmatization meant that co-occurrence relationships were defined with respect to lexemes, as opposed

to words. This increases the number of usable data points available from the corpus (the co-occurrence frequency for a pair of lexemes collapses together the counts for each combination of word forms mapping to the pair, boosting statistical reliability).

Co-occurrence vector parameters The relevant distributional information for the critical stimuli in each experiment was extracted from the BNC-spoken in the form of co-occurrence vectors. Vector representations were constructed by centering a "context window" over each occurrence of a target stimulus in the corpus, and recording the frequency with which each member of a pre-defined set of "context words" co-occurred with the target within the window. This set determines the dimensionality of the semantic space. We set the size of the context window to five words before and five words after the target word. The set of context words consisted of the 500 most frequent content words in the lemmatized corpus. Settings for the window size and dimensionality parameters were established empirically in a previous study of isolated word recognition (see McDonald & Shillcock, 2001b), and were not tuned to the current simulations.

ICE model parameters The ICE model employs two free parameters. Both parameters control the shape of the Dirichlet posterior distribution, and as such might be expected to affect the outcome. But we have tested the effects of different settings of these parameters, and it turns out that the pattern of our results is not sensitive to any but extreme variations

The first parameter determines how much weight should be given to prior information. Recall that the ICE model forms its probabilistically weighted hypotheses by integrating prior knowledge (derived from previous words in the context) with new data (the currently encountered word). The parameters $\underline{\alpha}$ of the Dirichlet prior distribution can be viewed as weights. For example, if the sum of the prior weights ($\underline{\alpha}_0$) is 1,000, and the results of 100 new "multinomial trials" are recorded, prior knowledge is ten times more important to the outcome than the newly arrived evidence.

After every update we scale the total prior weight (α_0) so that it is constant. This produces a straightforward discounting of old information, and is the simplest approach that we could find that has this biologically plausible property. We set the total prior weight parameter by maximizing the predictive probability of a small corpus (see McDonald, 2000, for details).

The second parameter is the scheme for determining the weight to be given to the incoming word. We could have given words weight in proportion to their frequency, but that would have given undue weight to frequent words. We therefore used a fixed size sample, setting the sample size parameter to 500. Although there are certainly other conceivable weighting schemes, this one is simple and easy to apply.

We do not claim that the ICE model is the optimal approach to inference, but merely that it uses aspects of Bayesian analysis that we think are justified. For example, although the iterative re-normalization of the prior weight represents a departure from a purely normative Bayesian model, its motivation is clear: to avoid "swamping" new evidence with prior knowledge. It is intuitively plausible (ignoring discoursal factors such as salience) that words occurring a long time ago have little effect on the prediction of the upcoming word's meaning.

Simulation Procedure

The procedure was identical for each of the three simulations described below. The ICE model was applied to the linguistic stimuli forming the materials for each original experiment, and only the distributional information associated with each stimulus in the corpus was available to the model. Each experiment used a design where the designated target word was presented immediately following a single-word (Simulation 1) or multiple-word (Simulations 2 and 3) context. Observations were always paired: the target word was held constant and the context was varied in order that any differences in the simulated processing of the target word could be attributed solely to the context manipulation.

For each item – the target word preceded by a semantically related (Simulations 1 and 2) or constraining (Simulation 3) context – the first application of the model was to the distribution derived from relative frequency information (the representation of the null context), and then successive words in the context were integrated using Bayes' Law, resulting in the posterior distribution. The fit between the "expected" meaning and "actual" meaning was estimated as the relative entropy between the posterior distribution and the target word's empirical distribution. The resulting quantity – in bit units – is the amount of information conveyed by the target word about its contexts of use, given the presence of a semantically related context. This procedure was then repeated in order to compute the information conveyed by the same target word when preceded by the second (unrelated or unconstraining) context paired with the target.

Simulation 1: Single-Word Priming

The first test of the ICE model was to simulate the results of Hodgson's (1991) single-word lexical priming study. We tested the hypothesis that a minimal priming context – a single word – would have a reliable effect on the amount of information conveyed by the target word, and that this effect would pattern with the human behavioral data. Specifically, we predicted that a related prime word (such as value) would reduce the amount of information conveyed by a target word (like worth) about its contexts of use, compared with an unrelated prime (such as tolerate). The difference in ICE values resulting from the divergent influence of the related and unrelated prime words on the form of the posterior distribution was expected to correspond to the difference in processing effort (measured by lexical decision response times) reported by Hodgson (1991, Experiment 1).

Hodgson employed prime-target pairs representing a wide range of lexical relations: antonyms (e.g., enemy-friend), synonyms (e.g., dread-fear), conceptual associates (e.g., teacher-class), phrasal associates (e.g., mountain-range), category co-ordinates (e.g., coffee-milk) and superordinate-subordinates (e.g., travel-drive). Using lexical decision latency as the dependent variable, Hodgson found equivalent priming

effects for all six types of lexical relation, indicating that priming was not restricted to particular types of prime-target relation, such as the category member stimuli employed by the majority of semantic priming studies. His materials represent a good test of the generality of the ICE model; if ICE succeeds in explaining the context effects obtained with prime-target pairs in diverse relationships using distributional information only, this would be good evidence that the human lexical processor is also able to exploit this source of information during word recognition.

Method

From the 144 original prime-target pairs listed in Hodgson (1991, Appendix), 48 were removed because either the prime or the target word (or both) had a lexeme frequency of less than 25 occurrences in the BNC-spoken. The reliability of co-occurrence vector representations decreases with word frequency (McDonald & Shillcock, 2001a), making it preferable to refrain from collecting statistics for low-frequency words than to compromise the generalizability of the modeling results. That is, comparable results would not be expected if using other similarly-sized corpora. The number of items remaining in each Lexical Relation condition after frequency thresholding is displayed in Table 1.

The ICE value for each Related prime-target combination was calculated using the model parameter settings detailed earlier. The corresponding value for each Unrelated item was computed as the mean of the ICE values for the target word paired with each of the other primes in the Lexical Relation condition.[4] For example, each Unrelated datapoint in the Antonym condition was computed as the mean of 15 ICE values.

[4] Because the unrelated primes corresponding to each target word were not supplied in Hodgson (1991), we used this technique to simulate the Unrelated context condition. An alternative would be to select a prime word at random from the other items in the same condition to serve as the unrelated prime; both methods give the same results.

Results and Discussion

We conducted a two-way analysis of variance on the simulated priming data generated by the ICE model. The factors were Lexical Relation (antonyms, synonyms, conceptual associates, phrasal associates, category co-ordinates, superordinate-subordinates) and Context (related, unrelated). ICE values for each cell of the design are presented in Table 1. (ICE values can be considered analogous to response times (RTs), the smaller the value, the shorter the RT).[5] As expected, there was a main effect of Context: collapsing across all types of Lexical Relation, target words conveyed significantly less information about their contexts of use when preceded by a related prime than when preceded by an unrelated prime: $F(1,90)=71.63$, MSE=0.0037, $p<0.001$. There was no main effect of Lexical Relation: $F(5,90)<1$, and importantly, no evidence for a Lexical Relation × Context interaction: $F(5,90)<1$. Separate ANOVAs conducted for each type of Relation showed consistent, reliable priming effects for all six relations: antonyms, $F(1,15)=14.82$, MSE=0.0050, $p<0.01$; synonyms, $F(1,10)=29.89$, MSE=0.0021, $p<0.01$; conceptual associates, $F(1,16)=13.45$, MSE=0.0047, $p<0.01$; phrasal associates, $F(1,19)=9.14$, MSE=0.0037, $p<0.01$; category co-ordinates, $F(1,17)=12.96$, MSE=0.0039, $p<0.01$; superordinate-subordinates, $F(1,13)=15.32$, MSE=0.0021, $p<0.01$.

[5] We systematically report simulation results in bits. A straightforward way to compare this measure to human response times (RTs) is the technique used by Plaut and Booth (2000). They regress their model condition means against the empirical condition means, and then use the regression equation to convert to RTs. Because we do not believe that prediction of meaning is the sole factor determining RT, and do not want to give the impression that the behavioural data could be fit perfectly with a better predictive model, we decided to present our simulation results using the unit appropriate to what is being measured.

Table 1

Mean ICE Values (bits) for Related and Unrelated Primes and Simulated Priming Effect (Difference) for Six Types of Lexical Relation

Lexical Relation	N	Context: Related	Context: Unrelated	Difference
Semantic				
Antonym	16	1.133	1.230	0.097
Synonym	11	0.673	0.736	0.063
Associate				
Conceptual	17	1.086	1.172	0.086
Phrasal	20	1.095	1.153	0.058
Category				
Coordinates	18	1.165	1.239	0.074
Super-subordinates	14	1.073	1.140	0.067

The simulated priming effects are clearly analogous to the human data. The results of the ICE simulation constitute a reasonable quantitative fit to the results reported by Hodgson (1991) using stimulus onset asynchronies (SOAs) varying from 83 to 500 ms. The overall pattern of semantic priming observed by Hodgson was replicated by the difference in the ICE value for a target word (worth) presented after a related prime (value), and the ICE value calculated for the same target word when presented after an unrelated prime (tolerate). For example, ICE was determined to be 0.628 bits for worth when preceded by value, and 0.833 bits for worth preceded by tolerate. The fact that value is a better cue to the meaning of worth than tolerate is reflected in these two ICE values; less information is conveyed by worth when following value.

The failure of the Context factor to interact with Lexical Relation indicates that distributional information present in the linguistic environment is sufficiently sensitive to the broad range of word-to-word relationships that produce priming in human subjects.[6] This finding expands upon successful semantic space model simulations of lexical priming between category co-ordinates (Lund, Burgess & Atchley, 1995; McDonald & Lowe, 1998).

In summary, Hodgson's (1991) results are compatible with an expectation-building view of contextual facilitation: if the context (the prime) allows the processor to create precise expectations about upcoming meaning, processing of a word that conveys that meaning (a related target) will be facilitated. Simulation 1 demonstrated that single-word lexical priming can be modeled as the influence of the local linguistic context on the quantity of information conveyed by a word about its contextual behavior.

In Simulation 2, we submit the ICE model to a more stringent test**:** the lexical priming situation where more than one prime word is presented before the target.

Simulation 2: Multiple Priming

The multiple priming paradigm – the procedure by which two or more lexical primes precede the target word – is a natural extension of the single-word priming task. Multiple priming can be seen as occupying the middle ground between the lexical priming and contextual constraint paradigms. In multiple priming experiments, the prime words are presented as unstructured lists, but in contextual constraint studies, whole

---

[6] How well does this result generalize to other sets of priming stimuli? The results of Simulation 1 are not specific to Hodgson's (1991) materials. We validated the ICE model's behavior using the large set of prime-target pairs employed by Becker and Joordens (1997). This set was of interest because it included adjectival, adverbial and nominal items. Target words reliably conveyed less information when preceded by a related prime word than by an unrelated (randomly re-paired) prime: $F(1,121)=127.58$, $MSE=0.0055$ $p<0.001$.

sentences are presented in their original order, and the usual cues to syntactic structure are present. Despite the fact that multiple primes do not form a syntactically coherent unit, research by Balota and Paul (1996) and Brodeur and Lupker (1994) has shown that two (or more) primes are better than one. In Simulation 2, we investigate the ability of the ICE model to explain this multiple-prime advantage.

Balota and Paul (1996) were interested in how multiple primes – construed as independent sources of spreading activation – influenced target word processing. Using two-word contexts, they separately manipulated the relatedness of each prime to the target word; this procedure allowed additive priming effects to be accurately measured. In their first experiment, Balota and Paul demonstrated that the multiple-prime advantage was additive: the facilitation obtained in the two-related-primes condition (RR) was equivalent to the sum of the facilitation for the one-related-prime conditions (UR and RU). (See Table 2 for sample stimuli). Because they found evidence for simple additivity using a range of prime presentation durations and both lexical decision and naming as response tasks (Balota & Paul, 1996, Experiments 1-5), the authors state that "… we believe that contextual constraints can produce simple additive influences on target processing." (p. 839). In terms of the ICE model, two related prime words would need to constrain the processor's expectations about the meaning of the target to a greater degree than a single related prime in order to simulate the multiple-prime advantage.

Balota and Paul also noted that previous multiple priming studies (e.g., Brodeur & Lupker, 1994; Klein, Briand, Smith & Smith-Lamothe, 1988) have confounded prime-target relatedness with relatedness between the prime words. In these studies, primes are members of the semantic category labeled by the target. For example, in the presentation <copper, bronze, metal> the two prime words are semantically related to each other, as well as to the target word metal. This leaves open the question of whether (employing the spreading activation metaphor) each prime independently activates the semantic category target. An alternative hypothesis would be that the first prime augments the activation level of the second prime, which in turn activates the target word to a greater

degree than it would have acting in isolation. Balota and Paul (1996, Experiment 1) address this concern by manipulating inter-prime relatedness; their design included a condition where both primes were related to a homographic target word, but unrelated to each other (e.g., <kidney, piano, organ>). For these stimuli, the first prime should not affect the activation level of the second, and hence the presence of additive priming could not be due to inter-prime relationships. The results of Balota and Paul's Experiment 1 indicated equal-sized additive priming effects for both types of target word (category label and homograph), leading the authors to argue that previous reports of the multiple-prime advantage (e.g., Brodeur & Lupker, 1994) cannot be the result of inter-prime facilitation.

The predictions of the ICE model do not distinguish between the independent activation and inter-prime facilitation accounts, as the model is not concerned with the mechanism by which priming takes place. However, it is not obvious what is to be expected. Modeling multiple priming requires two Bayesian updates. First, the expectations due to the initial prime are integrated into the (neutral) initial prior, then the result of this process is in turn treated as a prior and the expectations from the second prime are incorporated. Finally, the result is compared with the empirical contextual distribution, giving a measure of processing cost. Because the prior distribution is re-weighted with each iteration of the update rule, the order of presentation of the prime words might matter, as will the details of the distributions which represent the contextual information. Additivity, if observed, will be a consequence of contingent facts about these distributions, and does not follow from the specification of the model. We did think that we would find a multiple priming advantage under certain specific conditions: if both primes are drawn from the same semantic category (e.g., <copper, bronze>) the expectations that they evoke will be very similar, so the updates which they produce will be consistent, in the sense that they push an initially neutral prior expectation towards the same contextual distribution. We expected to see the multiple-prime advantage under these conditions. But if the primes are independent cues to different meanings of a homograph (as when the primes are <kidney, piano> and the target is organ) the first

update might push the expectation towards one region of semantic space whereas the second pushes the expectation towards another region (and, presumably, away from the first region, although maybe not by as much). We did not have any firm predictions for this condition.

The third finding of interest in Balota and Paul's experiments was the difference in the size of the priming effects between the UR and RU conditions. The contiguity of the related prime word to the target seemed to be important, since larger effects were obtained in the UR condition, over most of their experiments. Although the difference in mean RT between the RU and UR conditions rarely reached statistical significance in a single experiment, when the RT data from Experiments 1-5 was collapsed together this difference was highly reliable. Balota and Paul suggest several possible interpretations of this finding, including "disruption of priming by an intervening word" (p. 832). We expect the ICE model to capture this effect, since an intervening unrelated word will dilute expectations about the meaning of the upcoming target word.

In summary, our primary motivation for submitting Balota and Paul's (1996, Experiment 1) materials to a computational reanalysis was to assess the ICE model's ability to capture the pattern of contextual facilitation revealed using the multiple-priming paradigm. However, the simulation had three further goals; namely to shed light on Balota and Paul's claims regarding (1) the statistical equivalence of priming for category label and homograph target words, (2) the additivity of multiple primes, and (3) the influence of the temporal proximity of the related prime to the target word.

Method

The design was identical to that of Balota and Paul's Experiment 1. This was a 2×4 mixed factors design, with Type of Target (homograph, category label) as the between-items factor, and Prime Type (RR, UR, RU, UU) as the within-items factor.

Preparation of the lexical stimuli was very similar to the procedure carried out in Simulation 1. Inflected stimuli were first converted to their canonical forms, and items

containing target or related prime words that did not meet the 25-occurrence frequency threshold were removed. Unrelated prime words that failed to meet the frequency threshold were replaced with unrelated primes randomly chosen from the set of discarded items. From the 106 original homograph items, 69 could be used in the simulation. Out of the 94 original category stimuli, 39 met the frequency criterion. (See Table 2 for sample materials).

We computed ICE values for each target word when preceded by each of the four Prime Types. Model parameter settings were identical to those used in Simulation 1.

Results and Discussion

As in Simulation 1, facilitation was interpreted by a reduction in the amount of information conveyed by a target word about its contexts of use in one of the Related prime conditions (RR, RU and UR), compared with the UU (two-unrelated-primes) condition. Facilitation was apparent for all three Related conditions. The size of the context effect was 0.110 bits for the RR condition, 0.041 bits for the UR condition, and 0.079 bits for the RU condition. These differences in mean ICE value were verified by an analysis of variance, which revealed a main effect of Prime Type, $F(3,306)=40.53$, $MSE=0.0058$, $p<0.001$. There was no reliable effect of Target Type, $F(1,102)=2.26$, $MSE=1.421$, $p=0.135$, and no evidence for a Prime Type × Target Type interaction, $F(3,306)<1$. Table 2 displays the mean ICE values for the separated homograph and category label stimuli.

Table 2

Results of the Simulation of Balota and Paul (1996, Experiment 1), with Mean Lexical Decision Response Times (RT) and Amount of Priming (Priming)

	Example Stimuli			ICE	RT	Priming
Condition	Prime-1	Prime-2	Target	(bits)	(msec)	(msec)
Homograph targets						
RR	game	drama	play	0.895	601	34
UR	lip	drama	play	0.970	618	17
RU	game	tuna	play	0.932	630	5
UU	lip	tuna	play	1.011	635	
Category label targets						
RR	hate	rage	emotion	1.095	606	34
UR	author	rage	emotion	1.151	616	24
RU	hate	design	emotion	1.114	627	13
UU	author	design	emotion	1.193	640	

Note: R=related prime, U=unrelated prime.

The pattern of results was closely comparable to the human data. As expected, the strongest context effect was observed in the RR condition, which was larger than the effects in both the UR and RU conditions. This result replicates the multiple-prime advantage reported by Balota and Paul.

Additivity was also present in the simulation results. The sum of the facilitation obtained in the UR and RU conditions (0.041 + 0.079 = 0.120 bits) was nearly identical to the effect size obtained when both primes were related to the target (0.110 bits for the RR condition). The reduction in information conveyed by a target word preceded by two related primes is very close to the sum of the numbers resulting from separate presentations of each prime.

The results of the ICE simulation did not match the human data completely; specifically, the influence of the temporal proximity of the related prime was different to that reported by Balota and Paul. In our data, the context effect for the RU targets was larger than for the UR targets, whereas the pattern observed in human subjects was the opposite. This difference between the RU and UR conditions was statistically reliable: planned comparisons (with suitable alpha corrections) confirmed that all four conditions differed reliably from one other, at the α=0.05 level of significance. We investigated further, believing that the larger simulated priming effect for the RU condition was due to the differences between the Prime-1 words and the Prime-2 words. According to the ICE model, the independent (single-word) priming effect is larger for the Prime-1 words than the Prime-2 words: adopting the UU condition as the unrelated baseline, the effect size was 0.125 bits using only the Prime-1 set and 0.070 bits using the Prime-2 set. It appears that for the ICE model, Balota and Paul's set of Prime-1 words is substantially "better" than the second set.

Because we did not observe a Prime Type × Target Type interaction, we conclude that prime words did not have to be related to each other in order for the model to simulate the multiple-prime advantage. This is a noteworthy finding. Although the distributional information associated with pairs of prime words such as <kidney, piano> would clearly be expected to differ to a much greater extent than the information associated with pairs such as <copper, bronze> (corresponding to differences in semantic relatedness), the ICE model (like the human participants) still produced equivalently sized context effects. This behavior can be explained as follows: the distributional information associated with homograph targets like organ represents a conflation of the information associated with each distinct meaning. The posterior distribution resulting from application of the Bayesian update rule to two unrelated primes like kidney and piano can be seen as a weighted combination of two separate (and presumably very different) expectations. This conflated expectation matches the

contextual distribution associated with organ better than the expectation derived from either kidney or piano combined with an unrelated prime word.[7]

If one considers all possible modeling schemes that could capture multiple priming data, the one chosen here is reasonable. Data may exist which contradicts our assumption that priming can be modeled as an iterative updating of expectations. This process involves separate contributions from each prime word; because no part of the probabilistic model knows about more than one prime at the same time, interactions between prime words cannot be captured. Indeed, recent studies (Hutchison, Neely & Johnson, 2001; Neely, VerWys & Kahanm 1998; Pitzer & Dagenbach, 2001) have shown that repeating the prime word (e.g., <tale, tale, story>) can in certain circumstances reduce (or even eliminate) the semantic priming effect. This finding would appear to be at odds with both Balota and Paul's (1996) findings and the ICE model's predictions. Intuitively, expectations about the meaning of the target word should be more precise if the two prime words are related to each other or the same prime word occurs twice, leading to a larger predicted priming effect (which was

[7] ICE (and other models based on distributional statistics) employs a single vector to represent the distinct meanings of ambiguous words such as organ. Although conflating the distributional properties of the different usages of an ambiguous or polysemous word into a single representation is problematic for certain tasks in natural language processing, here this property is crucial – for kidney and piano to both prime organ, it is essential that the representation of organ be simultaneously similar to both prime words. However, at present the ICE model has to ignore the possible disruptive impact that incorporating an ambiguous context word into the prior has on the formation of expectations, while on-line contextual disambiguation is a indisputable feature of human comprehension.

obtained with the <copper, bronze, metal> stimuli by Balota and Paul).[8] However, by manipulating task parameters Hutchison et al. (2001) infer that prime repetition must influence strategically-mediated priming mechanisms, and so conclude that the repetition effect does not constitute evidence against automatic priming in the lexicon.

The simulation results are also compatible with the expectation-building hypothesis. If one attributes the basic single-word priming effect to the ability of a semantically related prime word to serve as a useful cue to the meaning of the target word, then the multiple-prime advantage reflects the greater degree of precision provided by two related primes compared with one. The statistically equivalent priming for the category label and homograph stimuli indicates that the semantic expectations formed from the sequential presentation of unrelated prime words like kidney and piano are as useful for predicting the meaning of the target word organ as presentation of related primes like <copper, bronze> are for predicting the meaning of metal. Even though prime words such as kidney and piano are not obviously related to each other, both words contribute to the formation of expectations about the meaning of the target word, facilitating the processing of organ.

In summary, the ICE model captures important features of the pattern of contextual facilitation obtained by Balota and Paul in their multiple-priming study. Although we did

---

[8] Applying the ICE model to Hutchison et al.'s (2001) materials confirms this prediction. A larger (though underadditive) simulated priming effect was achieved for the repeated prime condition (e.g., <tale, tale, story>) compared with the nonrepeated prime condition (e.g., <cope, tale, story>). An anonymous referee has suggested that such behavior might be an artifact of the Bayesian update rule, which would imply that the additive priming obtained in Simulation 2 was due to the re-normalization of the prior weight with each iteration, not the ability of the distributional properties of the primes to independently predict those of the target. However, applying the ICE model to the Hutchison et al. stimuli without holding the prior weight constant with each iteration also resulted in a repeated-prime advantage.

not observe the effect of temporal proximity, the multiple-prime advantage and simple additivity were present in the simulation.

In the next Simulation, we further test the hypothesis that the words making up the prime do not necessarily need to be related to each other for a priming effect to be observed, by modeling the contextual constraint effect.

Simulation 3: Contextual Constraint

Employing more ecologically valid stimuli than used in the standard lexical priming paradigm – grammatical sentences as opposed to sequentially presented individual words – sentence priming and contextual constraint studies have demonstrated that lexical processing effort is influenced by the properties of the linguistic context.

Contextual constraint has a robust effect on the eye movements that people make while reading silently. Eye-tracking technology allows an accurate temporal record to be made of the on-line processing of natural language, and thus seems ideal for testing word-by-word predictions about the effects of contextual constraint on natural reading processes. Contextually constrained words are fixated for less time and are skipped more often than words that are not constrained by the semantic context (e.g., Altarriba, Kroll, Sholl & Rayner, 1996; Rayner & Well, 1996; Schustack, Ehrlich & Rayner, 1987).

Differences in eye movement behavior can be seen even with constraint manipulations as small as the replacement of a single word. Schustack, Ehrlich and Rayner (1987) found that the probability of fixating on a target noun (e.g., floor) was significantly less when it was preceded by a "semantically restrictive" (i.e., constraining) verb (sweep). The control was the same word, preceded by a less restrictive verb (clean), in the same paragraph context.

To the extent that the crucial underlying mechanisms are the same as those of single-word and multiple-word priming, we expect that the ICE model will simulate the contextual constraint effect. Although the model is ignorant of the syntactic

dependencies holding between words in a sentence fragment, and therefore is unable to take specific account of syntactic constraints, it may be that the content words in the context, viewed as a simple sequence of words, are adequate predictors. On the other hand, if significantly different mechanisms are involved, there is no reason to expect that the ICE model will produce the right pattern of results.

The aim of Simulation 3 was to further test the ICE model, using pairs of multi-word sentence contexts that varied in how strongly they constrained the same target word. The materials from Altarriba et al.'s (1996, Experiment 1) eye movement study were ideally suited; these items produced reliable effects of constraint on skip probability and first fixation duration.[9] (Contextual constraint was determined using a cloze procedure [Taylor, 1953]). Altarriba et al. also manipulated the target word frequency, finding evidence for a frequency effect on gaze duration, but no interaction between the two variables. In this Simulation, we also examine the influences of both the Constraint and Frequency factors, treating ICE value as the dependent variable. We predicted independent effects of each factor on ICE values. A frequency effect was anticipated because, in general, high frequency words tend to convey less information about their contexts of occurrence than low frequency words (McDonald & Shillcock, 2001b).

Method

Inflected word forms in the material set were first converted to their canonical (lexeme) forms. The original stimuli consisted of 32 high-frequency and 32 low-frequency target words, with a high- and low-constraint sentence context created for

[9] It should be noted that Altarriba et al.'s (1996) subjects were Spanish-English bilinguals, and thus were processing sentences in their second language. The fact that their eye movement data were consistent with published data elicited from monolingual English-speaking subjects indicates that Altarriba et al.'s findings are representative of results obtainable from monolinguals.

each target. Four of the low-frequency items were removed because their target words failed to meet the lexeme frequency threshold of 25 occurrences in the BNC-spoken. The high-frequency targets had a mean lexeme frequency of 172 per million, compared with 17 per million for the low-frequency targets. The sentence contexts for the high-frequency target word teeth are displayed in (2), and the contexts for the low-frequency target thief are given in (3):

(2) a. The dentist told me to brush my teeth after every meal. (high-constraint)

b. He lost three teeth and had a black eye after the fight. (low-constraint)

(3) a. The robbery was committed by a thief who was known for his skill in safe-cracking. (high-constraint)

b. He warned us that the thief had escaped from prison on Wednesday. (low-constraint)

Target words and content words in the context preceding each target were first replaced with their canonical forms, if necessary. ICE values for each target word were computed adopting the same parameter settings used in Simulations 1-3. Note that we applied the model to only the content words in the context preceding the target word. The mean number of context words contributing to the Bayesian updating procedure was 3.2 words for the high-constraint contexts and 2.3 words for the low-constraint contexts.

Results and Discussion

Figure 2 displays the results of the simulation. The pattern of ICE values closely replicated the pattern of eye movement measurements reported by Altarriba et al.[10] ICE

[10] Altarriba et al. (1996, Experiment 2) confirmed the results of their eye movement study by using the same materials in a rapid serial visual presentation paradigm, although the main effect of frequency failed to reach significance.

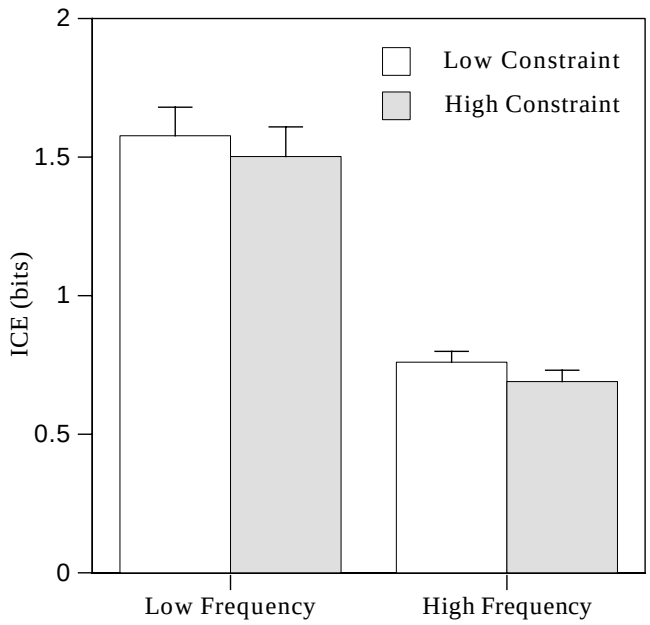

Figure 2. Computed ICE values in the simulation of Altarriba, Kroll, Sholl & Rayner (1996, Experiment 1) as a function of contextual constraint and word frequency (with standard errors).

values were smaller for high-constraint contexts than for low-constraint contexts, and were also smaller for high-frequency target words compared with low-frequency targets. An analysis of variance indicated reliable main effects of both Constraint and Frequency: $F(1,58)=22.15$, $MSE=0.0070$, $p<0.001$, and $F(1,58)=58.53$, $MSE=0.3383$, $p<0.001$, respectively. There was no interaction: $F(1,58)<1$. Separate ANOVAs on the high-frequency and low-frequency items also revealed significant Constraint effects: $F(1,31)=17.44$, $MSE=0.0044$, $p<0.001$, and $F(1,27)=7.83$, $MSE=0.0099$, $p<0.001$, respectively.

The simulation replicated the effects of constraint on skip probability and first fixation duration reported by Altarriba et al. Also, as expected, we found that the amount of information conveyed by high frequency target words was significantly less than the amount conveyed by low frequency words. By measuring gaze durations, the human studies found a corresponding frequency effect.

Making, as it does, no call on human subjects, the ICE model permits cheap and easy exploration of properties of the context that may be responsible for the constraint effect. Recall that the model incorporates parameters for weighting the contribution of new evidence to prior knowledge. The most recently encountered words contribute more weight to the final form of the posterior distribution – and therefore the computed ICE value – than earlier presented words. Examples (2) and (3) are representative of Altarriba et al.'s stimuli; note that the low-constraint contexts tended to have fewer content words preceding the target than the corresponding high-constraint contexts. Is it surprising, then, that the high-constraint contexts lead to more refined expectations? If, as suggested in Simulation 2, the more words in the context semantically related to the target, the larger the context effect, than the simulated constraint effect may represent more of a "weight of evidence" effect than a difference in the precision of expectations. We can address this concern by equating the number of context words considered by the ICE model in each constraint condition.

We wanted to know whether the present contextual constraint effect could be attributed entirely to the immediately preceding word. For example, is the verb brush – not the sentence fragment "the dentist told me to brush" – solely responsible for the facilitation observed on teeth? We re-ran the ICE simulation by truncating all stimuli to single-word contexts (the first content word to the left of the target) and found a small, though significant constraint effect of 0.021 bits: $F(1,58)=5.41$, MSE=0.0024, $p<0.05$. It indeed appears that a large portion of the predictability of the meaning of the target word, and thus the effect of contextual constraint, arises from the immediately preceding word. However, our analysis also revealed that this effect can be attributed mainly to the subset of high-frequency target words: $F(1,31)=8.40$, MSE=0.0023, $p<0.01$; there was no reliable constraint effect for the low-frequency targets: $F(1,27)<1$. It is not clear whether the absence of the single-word constraint effect for the low-frequency items is due to the nature of the contexts constructed for these items or to properties of the target words themselves. Pre-tests conducted by Altarriba et al. ensured that the cloze values

were equivalent for the high- and low-frequency items, so it seems unlikely that the high-frequency contexts were more constraining.

General Discussion

The ICE model simulations have demonstrated that the effects of three representative types of semantic context manipulation can be replicated using a simple model that relies solely on the distributional properties associated with words in real language, as recorded in a large corpus of spoken English. We have provided accounts of single-word lexical priming, multiple-priming, and the effects of contextual constraint that do not depend on detailed hypotheses about processing architectures or mechanisms. The current paper's primary contribution to the literature on semantic priming centers on the presentation and evaluation of a highly constrained model that draws upon distributional statistical information – information about lexeme co-occurrence – available from the linguistic environment.

We expect that the human language processor has better means at its disposal than the simple sequence-of-words model which we have used, but our point is that a substantial proportion of the interesting regularities found in human experiments can be explained by a simple instantiation of this central probabilistic idea. There can be no other reason for the success of the ICE model than a parallelism between distributional statistics and essential elements of word meaning. The ICE model is also consistent with what is known about the incremental nature of on-line language comprehension (e.g., Altmann & Kamide, 1999; Sedivy, Tanenhaus, Chambers & Carlson, 1999), as it implements the construction of expectations about upcoming word meaning as an incremental process of integrating prior knowledge with current evidence.

In fleshing out the details, we proposed that lexical processing cost reflects the "fit" between expected meaning and actual meaning – the better the match, the less the cost to the processing mechanism. Mathematically, this idea reduces to the use of relative entropy as a measure of the difference between distributions; functionally, it corresponds to the hypothesis that the human language processor uses a strategy which

is sensitive to the costs and benefits of occasionally making risky assumptions about what will probably appear next in the input stream.

The principal strength of the ICE model is that it provides a unified explanation of a range of semantic context effects, and that it accommodates these in a single framework. Psycholinguistic research has employed a variety of measurement techniques in order to establish the circumstances by which a diverse set of contextual influences are able to affect the effort of lexical processing. Our simulations add another technique: one that does not depend on the availability of human subjects.

Distributional Information and Meaning

The ICE simulations have also provided a demonstration of a fact which is easily overlooked, namely that the effects of priming and contextual constraint, while conventionally assumed to be "semantic", should perhaps be more neutrally characterized as "distributional". This is made very obvious when one notes that the only information available to the ICE model about the meaning of a word is what is latent in its pattern of use. In the absence of a detailed theory of semantic representation it is hard to see how the "semantic" and the "distributional" could ever be teased apart.

Without further stipulations, the model also captures the well-established effect of corpus frequency on lexical processing effort; Simulation 3 replicated the effect of frequency on gaze duration observed by Altarriba et al. (1996). ICE values are naturally correlated with word frequency – the less frequent a word, the more information it tends to carry about its contexts of occurrence (McDonald & Shillcock, 2001b) – and thus predicts the attested influence that frequency has on lexical processing effort. ICE is the theoretically more interesting explanatory variable, because it makes predictions which go beyond those of frequency alone.

Whilst distributional information has recently been recognized for its putative contribution to the acquisition of syntactic knowledge by children (Redington, Chater & Finch, 1998), and to the segmentation of the speech stream into words by infants (Brent

& Cartwright, 1996; Saffran, Aslin & Newport, 1996), there are few quantitative studies of its role in the (more or less) steady state of adult language comprehension and production. The ICE simulations provide support for a broad class of theories in which low-level information about patterns of lexical co-occurrence plays a major role. However, these simulations do not distinguish between theories in which this experience is acquired mainly in early childhood and those in which later language exposure leads to continuous adjustments in the policies adopted by the human language processor.

We can rule out a simpler model of the influence of word co-occurrence probabilities on lexical processing behavior. Distributional information in the form of the joint probability of two words co-occurring together in a corpus has been suggested by McKoon and Ratcliff (1992) to predict lexical priming. In the compound cue model of priming (Dosher & Rosedale, 1989; McKoon & Ratcliff, 1992; Ratcliff & McKoon, 1988), the priming effect (in the lexical decision task, at least) depends on the familiarity of the prime-target combination; the more familiar the combination, the quicker a word response can be made to the target word. McKoon and Ratcliff (1992) propose that cue familiarity can be measured by the corpus co-occurrence probability of the prime-target pair. Although they found that stimuli selected solely according to a measure of corpus co-occurrence gave rise to priming in a single-word priming task (McKoon & Ratcliff, 1992, Experiment 3), Williams (1994, Experiment 1) showed that high co-occurrence frequency was insufficient to trigger priming in a visual masked paradigm; semantic relatedness between prime and target was also required. Co-occurrence probability is naturally confounded with several other variables, including semantic similarity and normative association strength, making it difficult to attribute priming effects to a single factor.

We emphasize that the ICE model goes beyond simple co-occurrence. Although the ICE model ultimately exploits exactly this type of information, the expectations formed by the model concerns the distributional properties of the upcoming word – the approximate region of the 500-dimensional semantic space in which the upcoming word is expected to be fall – not its identity. For example, upon encountering value, the model

forms expectations for worth because the distributional properties of value in real language are similar to those of worth, not because value and worth often co-occur together in a corpus or because worth is highly predictable (in a corpus) given the presence of value. We believe that for certain lexical relations, the ICE model subsumes the notion of linear co-occurrence. For example, Hodgson's (1991) phrasal associates (e.g., mountain, range) would presumably figure strongly on a measure of simple co-occurrence; the target word range is intuitively highly predictable given the presence of the prime mountain, and the conditional probability P(range|mountain) would be higher than the conditional probability for range given an unrelated word such as foreign. The fact that the attested priming effect for such pairs of words was simulated by a significant difference in their ICE values indicates that the ICE model provides a more general account than simple measures of joint or conditional co-occurrence probability.

Other high-dimensional models derived from lexical co-occurrence statistics have been proposed to account for processing phenomena such as semantic priming (e.g., Lund, Burgess & Atchley, 1995; McDonald & Lowe, 1998) and homograph disambiguation (Landauer & Dumais, 1997). These models explain context effects in terms of representational similarity – the degree of similarity between the co-occurrence vectors extracted for prime and target words is assumed to correspond to the amount of contextual facilitation. Although these models partially overlap with the ICE model in terms of explainable behavior, we believe that the ICE approach is superior for three reasons. First, ICE is formulated in probabilistic terms, allowing a solid Bayesian interpretation of the expectation-forming process as the construction of probabilistically weighted hypotheses. Second, consistent with what is known about on-line language comprehension, these hypotheses are constructed incrementally. Third, unlike existing co-occurrence-based models, the ICE model is capable of providing behavioral predictions for the effort involved in processing words presented both in isolation and in context (see McDonald, 2000 ;McDonald & Shillcock, 2001b).

Comparing ICE to Alternative Models of Priming

Although our goal was to develop a model of lexical processing that was as constrained as possible – not specifically a model of priming – it is worthwhile to briefly compare ICE to recent computational models of semantic priming. We focus on two of the most constrained models in the literature: Cree, McRae and McNorgan's (1999) attractor network (see also McRae et al., 1997) and Plaut and Booth's (2000) distributed neural network model.

Cree et al. (1999) present a series of connectionist simulations that succeed in capturing the basic semantic priming effect as well as priming differences due to degree of similarity, and demonstrate that featural overlap, not shared superordinate categories underlie semantic-similarity priming. Thus, Cree et al.'s simulations are consistent with a distributed view of semantic memory organization. The ICE model can also be seen as drawing upon an analogous conception of featural overlap. The relative entropy measure used to estimate the fit between expectations and meaning is in fact a measure of overlap, so the primary difference between ICE and Cree et al.'s network is the choice of the features. Whilst Cree et al. employed subject-generated features for their model's semantic representations, we have obtained our "features" from the data – words occurring in the immediate context of another word in the corpus. Although we consider the ICE model to be a functional account of expectation-formation and Cree et al. describe their network as "an instantiation of how word meaning is computed" (p. 409), it is apparent that at the computational level of description, the models are quite similar. To provide a fuller account of the human data, both models require mechanistic assumptions and extensions to deal with task-specific strategic effects and time-course differences seen with semantic priming.

The distributed network model presented by Plaut and Booth (2000) is a comprehensive account of a range of lexical priming phenomena. Plaut and Booth focus on modeling individual and developmental differences in priming with a single mechanism, and their model successfully captures both these and differences attributable

to the manipulation of the delay between prime and target presentation. As ICE is not a process model, it has little to say about the time course of lexical processing. The dependence of context effects on temporal factors may not be part of the domain of expectation-building, but rather be located in the processes concerned with if, when, and how expectations are used to aid comprehension. Although Plaut and Booth's model does not employ empirically-derived semantic representations (they employ random binary vectors), featural overlap is still the fundamental organizing principle. Their simulations reveal the importance of perceptual ability in lexical processing, and demonstrate that a distributed network can capture certain strategic influences on priming without the need for complex multiple mechanisms.

In summary, we see the key features of the ICE model as: (1) it is highly constrained yet captures detailed behavioral data; (2) its tight link between the nature of a word's representation and the environment in which it occurs; and (3) its principled, probabilistic interpretation of expectation-formation.

The Relationship between ICE and Mechanistic Theories

Can our simulation results say anything useful about the cognitive mechanisms that underlie lexical processing? For instance, the lexical decision task involves demanding hand-eye coordination and decision-making processes not required for normal reading; therefore there will be influences on lexical decision response latency and error rate that are specific to the demands of the task. What the ICE model offers is a way to investigate the properties of the environment where the processing task takes place without the need to explain the precise means by which the task is accomplished by the cognitive system.

Is the ICE model compatible with current models of contextual facilitation, such as spreading activation theory? Because it is situated at the computational level of explanation – modeling the function of the language processor – the ICE model does not discriminate between the different means by which the function may be cognitively

realized. One might presume that spreading activation, along with many other possible models of contextual facilitation, is compatible with the range of empirical data examined in the current paper. But is this the most parsimonious approach? In the case of Altarriba et al.'s (1996) contextual constraint experiment, the observed additive influences of constraint and frequency would normally suggest that different stages of processing were involved, meaning that a separate mechanism accounting for frequency effects is required above and beyond the contextual facilitation mechanism. As we have shown in Simulation 3, the ICE model provides a parsimonious explanation of this data.

Our approach is simple, and involves few tunable parameters, and so lends itself to exploratory work and to the generation of clear and testable hypotheses. It is straightforward, given a large corpus and a sufficiently precise working hypothesis, to create sets of stimulus materials that should produce context effects, and to test them using human participants. In this respect, spreading activation models are less convenient; in order to account for the wide range of lexical relations that trigger priming (as shown by Hodgson, 1991), a multitude of links between concepts are required. As more lexical relations are discovered to produce context effects, the more complex the network of links (and thus the model) becomes. This gives flexibility, but runs the risk of over-fitting the data. While spreading activation may have too many explanatory variables, our approach almost certainly has too few. A standard computational solution to such problems is to penalize complex structures. This balances a preference for simplicity with the pressure to fit the data, and is a natural extension of the Bayesian techniques advocated in the current paper, with the additional benefit of potentially allowing the controlled introduction of further explanatory variables.

In summary, the experimental evidence to date has failed to isolate a particular unique mechanism responsible for semantic context effects. Spreading activation is not excluded as a candidate, but in its most general form has very little explanatory power. Non-invasive experimental techniques such as the recording of event-related brain potentials, which eliminate the need for an extraneous task, should allow a clearer picture of the relevant cognitive mechanisms to emerge. More importantly, such techniques are

useful for clarifying what might be happening during normal comprehension. This is exactly what Brown, Hagoort and Chwilla (2000) have shown using an ERP version of the standard semantic priming paradigm, with and without an accompanying lexical decision response task. They found that semantic priming was reflected in the N400 component of the ERP waveform, independently of the requirement for subjects to make lexical decisions. The conclusion is that contextual facilitation, as observed in a variety of experimental paradigms, is simply what is going on anyway. Our ICE simulations have demonstrated that the distributional properties of the linguistic environment can explain this natural aspect of ordinary language comprehension.

References

Altarriba, J., Kroll, J., Sholl, A. & Rayner, K. (1996). The influence of lexical and conceptual constraints on reading mixed-language sentences: Evidence from eye fixations and naming times. Memory & Cognition, 24, 477-492.

Altmann, G. T. M. & Kamide, Y. (1999). Incremental interpretation at verbs: restricting the domain of subsequent reference. Cognition, 73, 247-264.

Anderson, J. R. (1983). The architecture of cognition. Cambridge, MA: Harvard University Press.

Anderson, J. R. (1990). The adaptive character of thought. Hillsdale, NJ: Erlbaum.

Balota, D. A., Pollatsek, A. & Rayner, K. (1985). The interaction of contextual constraints and parafoveal visual information in reading. Cognitive Psychology, 17, 364-390.

Becker, C. A. (1980). Semantic context effects in visual word recognition: an analysis of semantic strategies. Memory & Cognition, 8, 493-512.

Brent, M. R. & Cartwright, T. A. (1996). Distributional regularity and phonotactics are useful for early lexical acquisition. Cognition, 61, 93-125.

Brodeur, D. A. & Lupker, S. J. (1994). Investigating the effects of multiple primes. Psychological Research, 57, 1-14.

Brown, C. M., Hagoort, P. & Chwilla, D. J. (2000). An event-related brain potential analysis of visual word priming effects. Brain and Language, 72, 158-190.

Chater, N. & Oaksford, M. (1999). Ten years of the rational analysis of cognition. Trends in Cognitive Sciences, 3, 57-65.

Collins, E. M. & Loftus, E. F. (1975). A spreading activation theory of semantic processing. Psychological Review, 82, 407-428.

Cree, G. S., McRae, K. & McNorgan, C. (1999). An attractor model of lexical conceptual processing: Simulating semantic priming. Cognitive Science, 23, 371-414.

Dosher, B. A. & Rosedale, G. (1989). Integrated retrieval cues as a mechanism for priming in retrieval from memory. Journal of Experimental Psychology: General, 2, 191-211.

Duffy, S. A., Henderson, J. M. & Morris, R. K. (1989). The semantic facilitation of lexical access during sentence processing. Journal of Experimental Psychology: Learning, Memory and Cognition, 15, 791-801.

Gelman, A., Carlin, J. B., Stern, H. S. & Rubin, D. B. (1995). Bayesian data analysis. London: Chapman & Hall.

Hodgson, J. M. (1991). Informational constraints on pre-lexical priming. Language and Cognitive Processes, 6, 169-205.

Hutchison, K. A., Neely, J. H. & Johnson, J. D. (2001). With great expectations, can two "wrongs" prime a "right"? Journal of Experimental Psychology: Learning, Memory, and Cognition, 27, 1451-1463.

Keefe, D. E. & Neely, J. H. (1990). Semantic priming in the pronunciation task: The role of prospective prime-generated expectancies. Memory & Cognition, 18, 289-298.

Klein, R., Briand, K., Smith, L. & Smith-Lamothe, J. (1988). Does spreading activation summate? Psychological Research, 50, 50-54.

Landauer, T. K. & 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.

Lund, K. & Burgess, C. (1996). Producing high-dimensional semantic spaces from lexical co-occurrence. Behavior Research Methods, Instruments, & Computers, 28, 203-208.

Lund, K., Burgess, C. & Atchley, R. (1995). Semantic and associative priming in high-dimensional semantic space. In Proceedings of the 17th Annual Conference of the Cognitive Science Society (pp. 660-665). Mahwah, NJ: Erlbaum.

Lupker, S. (1984). Semantic priming without association: a second look. Journal of Verbal Learning and Verbal Behavior, 23, 709-733.

Marslen-Wilson, W. D. & Tyler, L. K. (1980). The temporal structure of spoken language understanding. Cognition, 8, 1-71.

Masson, M. E. J. (1986) Comprehension of rapidly presented sentences – the mind is quicker than the eye. Journal of Memory & Language, 25, 588-604.

McDonald, S. A. (2000). Environmental determinants of lexical processing effort. Unpublished PhD dissertation, University of Edinburgh.

McDonald, S. & Lowe, W. (1998). Modelling functional priming and the associative boost. In Proceedings of the 20th Annual Conference of the Cognitive Science Society (pp. 667-680). Mahwah, NJ: Erlbaum.

McDonald, S. & Shillcock, R. (2001a). Contextual Distinctiveness: a new lexical property computed from large corpora. Informatics Research Report EDI-INF-RR-0042, University of Edinburgh.

McDonald, S. A. & Shillcock, R. C. (2001b). Rethinking the word frequency effect: the neglected role of distributional information in lexical processing. Language and Speech, 44, 295-323.

McKoon, G. & Ratcliff, R. (1992). Spreading activation versus compound cue accounts of priming: Mediated priming revisited. Journal of Experimental Psychology: Learning, Memory, and Cognition, 18, 1155-1172.

McRae, K., de Sa, V. R. & Seidenberg, M. S. (1997). On the nature and scope of featural representations of word meaning. Journal of Experimental Psychology: General, 126, 99-130.

Meyer, D. & Schvaneveldt, R. (1971). Facilitation in recognizing pairs of words: Evidence of a dependence between retrieval operations. Journal of Experimental Psychology, 90, 227-234.

Morris, R. K. (1994). Lexical and message-level sentence context effects on fixation times in reading. Journal of Experimental Psychology: Learning, Memory, and Cognition, 20, 92-103.

Moss, H. E., Ostrin, R. K., Tyler, L. K. & Marslen-Wilson, W. D. (1995). Accessing different types of lexical semantic information: Evidence from priming. Journal of Experimental Psychology: Learning, Memory, and Cognition, 21, 863-883.

Muckel, S., Scheepers, C., Crocker, M. & Müller, K. (2002). Anticipating German particle verb meanings: Effects of lexical frequency and plausibility. Paper presented at the Eighth Conference on Architectures and Mechanisms for Language Processing (AMLaP-2002). Tenerife, Spain.

Nebes, R. D., Boller, F. & Holland, A. (1986). Use of semantic context by patients with Alzheimer's Disease. Psychology and Aging, 1, 261-269.

Neely, J. H. (1991). Semantic priming effects in visual word recognition: a selective review of current findings and theories. In D. Besner & G. W. Humphrey (Eds.) Basic processes in reading: Visual word recognition (pp. 264-336). Hillsdale, NJ: Erlbaum.

Neely, J. H., VerWys, C. A. & Kahan, T. A. (1998). Reading "glasses" will prime "vision," but reading a pair of "glasses" will not. Memory & Cognition, 26, 34-39.

Oaksford, M. & Chater, N. (1998). An introduction to rational models of cognition. In M. Oaksford & N. Chater (Eds.) Rational Models of Cognition (pp. 1-18). Oxford: OUP.

Pitzer, K. D. & Dagenbach, D. (2001). A constraint on eliminating semantic priming by repeating a prime. American Journal of Psychology, 114, 43-53.

Plaut, D. C. (1995). Semantic and associative priming in a distributed attracter network. Proceedings of the 17th Annual Conference of the Cognitive Science Society (pp. 37-42). Hillsdale, NJ: Erlbaum.

Plaut, D. C. & Booth, J. R. (2000). Individual and developmental differences in semantic priming: Empirical and computational support for a single mechanism account of lexical processing, Psychological Review, 107, 786-823

Plaut, D. C. & Shallice, T. (1994). Connectionist modelling in cognitive neuropsychology: a case study. Hove, England: Erlbaum.

Ratcliff, R. & McKoon, G. (1988). A retrieval theory of priming in memory. Psychological Review, 95, 385-408.

Rayner, K. & Well, A. D. (1996). Effects of contextual constraint on eye movements in reading: a further examination. Psychonomic Bulletin & Review, 3, 504-509.

Redington, M., Chater, N. & Finch, S. (1998). Distributional information: a powerful cue for acquiring syntactic categories. Cognitive Science, 22, 425-469.

Saffran, J. R. Aslin, R. N. & Newport, E. L. (1996). Statistical learning by 8-month old infants. Science, 274, 1926-1928

Schustack, M. W., Ehrlich, S. F. & Rayner, K. (1987). Local and global sources of contextual facilitation in reading. Journal of Memory & Language, 26, 322-340.

Schwanenflugel, P. J. & LaCount, K. L. (1988). Semantic relatedness and the scope of facilitation for upcoming words in sentences. Journal of Experimental Psychology: Learning, Memory, and Cognition, 14, 344-354.

Schwanenflugel, P. J. & Shoben, E. J. (1985). The influence of sentence constraint on the scope of facilitation for upcoming words. Journal of Memory and Language, 24, 232-252.

Sedivy, J. C., Tanenhaus, M. K., Chambers, C. G. & Carlson, G. N. (1999). Achieving incremental semantic interpretation through contextual representation. Cognition, 71, 109-147.

Sharkey, N. E., & Mitchell, D. C. (1985). Word recognition in a functional context. Journal of Memory and Language, 24, 253-270.

Smith, E. E. & Medin, D. L. (1981). Categories and concepts. Cambridge, MA: Harvard University Press.

Tanenhaus, M. K. & Lucas, M. M. (1987). Context effects in lexical processing. Cognition, 25, 213-234.

Taylor, W. L. (1953). "Cloze Procedure": a new tool for measuring readability. Journalism Quarterly, 30, 415-433.

Van Berkum, J. J. A., Hagoort, P. & Brown, C. M. (1999). Semantic integration in sentence and discourse: Evidence from the N400. Journal of Cognitive Neuroscience, 11, 657-671.

Van Petten, C., Coulson, S., Rubin, S., Plante, E. & Parks, M. (1999). Time course of word identification and semantic integration in spoken language. Journal of Experimental Psychology: Learning, Memory, and Cognition, 25, 394-417.

Williams, J. N. (1996). Is automatic priming semantic? European Journal of Cognitive Psychology, 8, 113-161.

Wittengenstein, L. (1958). Philosophical investigations. Oxford: Blackwell.

Author Note

Scott A. McDonald, Department of Psychology, University of Edinburgh.

Chris Brew, Center for Cognitive Science, The Ohio State University, 20C Page Hall, 1810 College Road, Columbus, Ohio, USA

Correspondence concerning this manuscript should be addressed to Scott McDonald, Department of Psychology, University of Edinburgh, 7 George Square, Edinburgh EH8 9JZ, or by electronic mail to Scott.McDonald@ed.ac.uk.