A distributional model of semantic context effects in lexical processing

Scott McDonald
Department of Psychology
and
Institute for Adaptive and Neural Computation
University of Edinburgh
Edinburgh, Scotland, UK

Chris Brew
Department of Linguistics
and
Center for Cognitive Science
The Ohio State University
Columbus, Ohio
USA

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 computational model of contextual facilitation based on word co-occurrence vectors, and empirically validate the model through simulation of three representative types of context manipulation: single word priming, multiple-priming and contextual constraint. The aim of our study is to find out whether special-purpose mechanisms are necessary in order to capture the pattern of the experimental results.

1 Introduction

In psycholinguistics, lexical access is the process of retrieving a word from the mental lexicon using perceptual and contextual information. In everyday life, the point of this process is to facilitate communication. Many different experimental methodologies have been brought to bear on the study of this process, including visual and auditory lexical decision tasks (e.g., Meyer & Schvaneveldt, 1971; Moss, Ostrin, Tyler & Marslen-Wilson, 1995), event-related brain potentials (e.g., Brown, Hagoort & Chwilla, 2000), and the recording of eye movements during normal reading. 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).

Because these effects are robust and apparently automatic, researchers often seek explanations in terms of low-level mechanisms such as spreading activation, 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. 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 the literature. It is not necessary to invoke distinct mechanisms for the different priming settings. Furthermore, we did not need to vary the independently tunable parameters of our algorithm in order to obtain our results. The same model has been used in simulations of eye movement behavior during reading (McDonald, 2000) and event-related potentials recorded from the brain (McDonald & Brew, 2001).

2 Distributional models

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, which we call the ICE model (for Incremental Construction of semantic Expectations), is of this type. Specifically, it maintains a vector of probabilities as its representation of the current best guess about the likely location in semantic space of the upcoming word. We use the semantic space defined by the 500 most frequent content words of the spoken portion of the British National Corpus (BNC-spoken).

When a 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.

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 normally ascribed to a multitude of mechanisms.

2.1 The ICE model

We use a vector-based representation of the "best-guess" hypothesis about context.

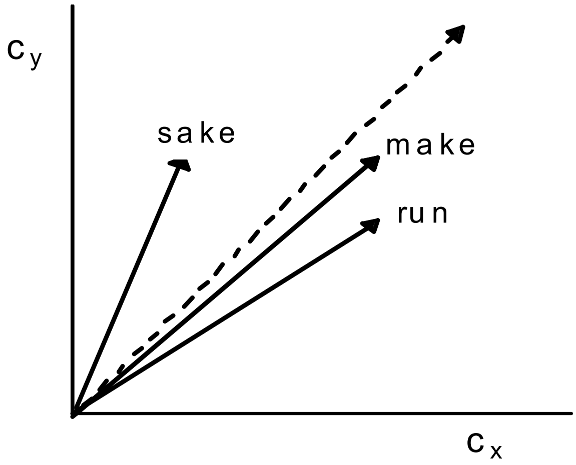

Figure 1. Distributional representations of "sake","make","run" and the null context.

The vector representations in Figure 1 encode the number of times the window five words to either side of the target word is discovered to include each of two context words $\mathbf{c_x}$ and $\mathbf{c_y}$ (Figure 1 shows the semantic space as having two dimensions, rather than the 500 actually used in our simulations.) The representations for the real words are formed by examining the distribution of context words in the neighborhood of these target words, while the representation of the null context is derived from the distribution of context words over the corpus as a whole. "Sake" is shown as having a distributional representation far from that of the null context, while "make" and "run" are relatively close. Therefore we predict that it will be harder to process "sake" in the null context than it is to process the other words.

The account of priming: in single word priming, we need two moves within the semantic space. Consider the case of the word "metal" being primed by the word "bronze", as shown in Figure 2. The first step moves the system's hypothesis away from the null context. The resulting intermediate position is shown as the diagonal of the quadrilateral linking "metal" to the origin and the null context.

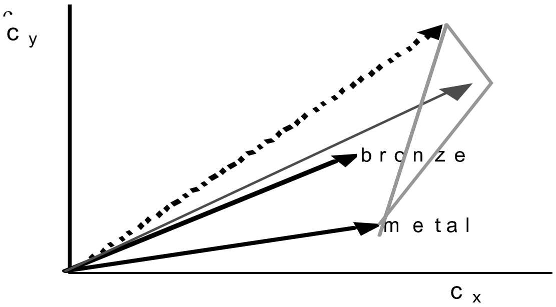

Figure 2. A distributional account of priming.

In the second step, the system needs to move from the intermediate position to the final position, which is the vector representation of the target word. The relative entropy between the intermediate position and the target distribution is our simulation of the effort expended by the lexical processor in understanding the word.

The model: Our model is Bayesian; the vectors shown in the diagrams are summaries of its ongoing estimates of the underlying high-dimensional probability distribution that gives rise to the observed distribution of co-occurring context words. This licenses the use of relative entropy, which we employ as the primary dependent variable in our simulations of semantic context effects. Because every distribution that we consider involves a contribution from a highly unspecific distribution associated with the null context, there are no zeroes in the distributions, and relative entropy can be used directly, with no need for smoothing.

The distributions: We can simulate meanings using multinomials – computing relative entropy by comparing entries in the 500-dimensional vectors associated with the context words – but to model the dynamic processes involved in semantic priming we need to represent more than just the maximum of the likelihood. We also want to model the extent to which the lexical processor is committed to the hypothesis that the target will be found in the location that we expect.

For reasons of simplicity we prefer distributions that have convenient analytical properties and concise parametric representations. One such is the *Dirichlet distribution*, which is widely used in Bayesian statistics (Gelman, Carlin, Stern & Rubin, 1995). We begin with prior information expressed in the form of a Dirichlet, then update it with data drawn from a multinomial. The resulting posterior distribution is also a Dirichlet, albeit one whose parameters have been adjusted to better fit the recently observed data. This closure property (known in the statistical literature as conjugacy) is crucial to our application, since it allows us to model both prior and posterior hypotheses in the same way.

The difference between the Dirichlet and the multinomial is that the latter is parameterized by a vector of probabilities, subject to the constraint that the sum must be zero, while the Dirichlet is specified by a vector of arbitrary real-valued weights, subject to no such constraint. It represents both a direction in semantic space and the number of "virtual samples" on which the estimate of that direction is based. It can therefore be used in priming simulations to represent both the current best guess about the upcoming word's position in semantic space and the strength with which this belief is held.

We need to decide how the balance will be struck between the prior and the incoming new word, and we need to implement some discounting strategy to prevent the weight given to the prior from increasing without limit and overwhelming the incoming data.

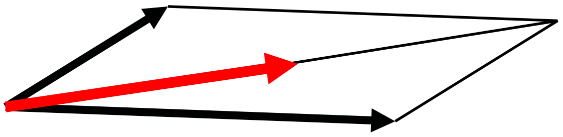

Figure 3: Discounting

To avoid this, we first add together the vectors corresponding to the two words, then shrink the result, as shown in Figure 3. The sum of the two word vectors is the full diagonal of the quadrilateral, while the shrunken version is just the bold part of the diagonal.

Model parameters: The ICE model has two free parameters. 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). For example, if the sum of the prior weights is 1000, and the results of 100 new “multinomial trials” are recorded, prior knowledge is deemed ten times more important to the outcome than the newly arrived evidence.

After every update we scale the total prior weight 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 to 2000 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. Thus, our model weights prior context as four times more significant than the incoming word. We have tested the sensitivity of our results to variations in this parameter, and the results are not significantly impacted by any but the largest changes. Although there are certainly other conceivable discounting schemes, this one is simple, robust, and easy to apply.

This tells us how to generate the vectors for the intermediate stage of the priming process, producing new positions in the semantic space. We compare this positions using relative entropy, just as if they were ordinary multinomials. Although it is only an approximation, it works well, as the results below demonstrate.

The account of multiple priming: with the Dirichlet-based simulation of priming in hand, the simulation of multiple priming is easy. We just need three steps instead of two.

Reaction time modeling: Our Bayesian measure is only one of the components that would be needed in a full mechanistic model of human reaction time (RT) behavior. To do justice to the richness of RT data, one would need to model not only the effects of informational context but also those of time pressure and experimental setup. Our model could be used, for example, to parameterize a diffusion model (Ratcliff & Smith, 2004).

3 Simulations

We used the same model settings for three experiments, of which two are reported here. The third is simulation of a contextual constraint study by Altarriba, Kroll, Sholl and Rayner (1996). This is described in a longer version of the present paper (McDonald & Brew, 2002).

3.1 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 relative entropy of a target word (like *worth*), 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 lexical decision response times reported by Hodgson (1991, Experiment 1).

Hodgson (1991) 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*). 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.

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, 2001), making it preferable to refrain from collecting statistics for low-frequency words. 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.[1] For example, each Unrelated datapoint in the Antonym condition was computed as the mean of 15 ICE values.

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 reaction times, the smaller the value, the shorter the RT). As expected, there was a main effect of Context: collapsing across all types of Lexical Relation, relative entropy was significantly less when the target is preceded by a related prime than when it is 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

As was the case for human subjects, Context did not interact with Lexical Relation. There is no evidence here for different mechanisms for the different types of word-to-word relations. We know that ICE is using nothing but distributional information, and it could be that human subjects are doing the same.

3.2 Simulation 2: multiple priming

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. 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 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 others has shown that two (or more) primes are better than one.

Balota and Paul 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, they 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

[1] 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.

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.

Table 1: Mean ICE Values (bits) for Related and Unrelated Primes and Simulated Priming Effect (Difference) for Six Types of Lexical Relation

Lexical		Context		
Relation	N	Related	Unrelated	Effect
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

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)

Condition	Prime-1	Prime-2	Target	ICE (bits)	RT (msec)	Priming (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.

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 simulated by a reduction in relative entropy 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.

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. The results of the ICE simulation did not match the human data completely; specifically, 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 $\alpha=0.05$ level of significance. We investigated further. Briefly, it appears that the discrepancy may be an artifact of the particular choice of experimental materials. The larger simulated priming effect for the RU condition was probably due to the differences between the Prime-1 words and the Prime-2 words.

4 Conclusions and future work

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. Because of the multiplicity of relevant linking relations evidenced by Hodgson, 1991, this would be harder to do in a spreading activation framework.

Another avenue for exploration is to use the combination of ICE and the refined lexical relations encoded in WordNet to create materials that would allow a larger scale replication of the results of Hodgson (1991). Such replication is independently desirable, since new reaction times would address the potential objection that we have unintentionally tuned our method to Hodgson's data. In the same vein, since our distributional methods provide a cheap and easy tool for exploratory studies, we intend to look more closely at the reasons for the discrepancies between our results and those of Balota and Paul (1996).

The present simulations show that a range of contextual effects can be subsumed under the same distributional mechanism, and that no task specific tuning of the parameters is necessary. Our model is computationally efficient and usable on a large scale to mine corpora for potentially interesting experimental materials.

Acknowledgements

Financial support for the first author's Ph.D studies came from the National Sciences and Engineering Research Council of Canada, the Overseas Research Student Awards Scheme, the Sir Ernest Cassels Education Trust, and the Institute for Adaptive and Neural Computation. The work of preparing this paper was funded in part by NSF Career Grant 0347799 to the second author. We are grateful to the Clippers discussion group at the Ohio State University for the opportunity to present and discuss this work from a computational perspective. Gail McKoon provided particularly useful feedback.

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.

Balota, D.A., & Paul, S.T. (1996). Summation of activation: Evidence from multiple primes that converge and diverge within semantic memory. Journal of Experimental Psychology: Learning, *Memory, and Cognition, 22*, 827-845.

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.

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

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.

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.

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

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

McDonald, S. & Brew, C. (2002). A distributional model of semantic context effects in lexical processing. *Cogprints.*

McDonald, S. & Brew, C. (2001). A rational analysis of semantic processing by the left cerebral hemisphere. *First Workshop on Cognitively Plausible Models of Semantic Processing (SEMPRO-2001)* , Edinburgh. July 31, 2001

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. A. & Shillcock, R. C. (2001). 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.

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.

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.

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.

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

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

Ratcliff, R. & Smith, P.L. (2004). A comparison of sequential sampling models for two-choice reaction time. *Psychological Review, 111*, 333-367

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