Division of Informatics, University of Edinburgh

Institute for Adaptive and Neural Computation
Institute for Communicating and Collaborative Systems

Contextual Distinctiveness: A New Lexical Property Computed from Large Corpora

by

Scott McDonald, Richard Shillcock

Informatics Research Report EDI-INF-RR-0042

Division of Informatics **July 2001**
http://www.informatics.ed.ac.uk/

Contextual Distinctiveness: A New Lexical Property Computed from Large Corpora

Scott McDonald*
Richard Shillcock*†

**Institute for Adaptive and Neural Computation*
Division of Informatics
†Department of Psychology

University of Edinburgh, Edinburgh, Scotland, UK

ABSTRACT

We describe the computational formulation of Contextual Distinctiveness (CD), a new lexical property derived from the distributional information present in natural language corpora. CD measures the quantity of information a word conveys about its contexts of use, which we demonstrate to be an interesting and objective indicator of the distributional differences between words. CD is computed from co-occurrence vector representations created using similar methodology to that of Lund and Burgess (1996) and Landauer and Dumais (1997), but provides a means to quantify between-word differences in contextual behavior. We establish the psychological relevance of CD to lexical processing behavior by showing that CD values are significantly correlated with published lexical decision and naming latencies obtained in an isolated word recognition task.

INTRODUCTION

The recent availability of large language corpora – multi-million word records of natural language output – has allowed large-scale computational modelling of a number of psycholinguistic phenomena. The adoption of corpus-based methods in psycholinguistic research has provoked new ways of thinking about, for example, normative association strength (Spence & Owens, 1990), the acquisition of syntactic categories (e.g. Redington, Chater & Finch, 1998), semantic and associative priming (Lund, Burgess & Atchley, 1995; Lund, Burgess & Audet, 1996; McDonald & Lowe, 1998), and vocabulary learning (Landauer & Dumais, 1997). Of particular interest is the ease with which aspects of a word's *environment* – its linguistic contexts of use – can be analysed, in order to estimate the semantic similarities and differences between words. A number of researchers have shown that simple lexical co-occurrence statistics contain much information about a word's meaning (e.g. Lund & Burgess, 1996), information which previously could only be compiled through collection of large numbers of human judgements. Corpus-based methods are attractive because information about a word's contexts of use can be easily and economically collected for a huge portion of the lexicon, tens of thousands of words – an insurmountable task for conventional methods relying on human intuitions.

Research activity has concentrated on the investigation of semantic space models – high-dimensional spaces where the positions along each axis corresponds to lexical co-occurrence frequencies extracted from a corpus of natural language. In this type of model, a word is represented as a vector, where the components of the vector are

labelled with other words (the *context words*), and the value of each vector component encodes the number of times the word of interest co-occurs with the component label, within a pre-defined *window* of words. Assuming *k* context words, a word *w* can be viewed as a point in *k*-dimensional space, and standard geometric distance and similarity measures can be applied to any two words. The co-occurrence vector for *w* can be considered to be a high-dimensional summary of its contextual behavior.

Semantic space models have been principally applied to the investigation of representational issues – how words are represented in semantic memory, and the processing implications for words with similar representations. Co-occurrence information offers a very different explanation for the ubiquitous lexical priming effect – the facilitation of a target word such as *coat* when preceded by a semantically related *prime* word such as *hat*, compared with an unrelated prime such as *pencil*. In semantic space models of lexical priming (Lund, Burgess & Atchley, 1995; Lund, Burgess & Audet, 1996; McDonald & Lowe, 1998) it is assumed that the difference in processing effort between the related and unrelated conditions is a direct reflection of representational similarity. Because the co-occurrence vectors for semantically related words are 'closer' in semantic space than unrelated words, less effort is required to process a particular target word when preceded by its related prime than when preceded by an unrelated word.

High-dimensional vector representations have also been used to represent the meaning of text units longer than a single word. Landauer and Dumais (1997) average together the vector representations for all content words in a unit of text (such as a sentence), and showed that the resulting centroid vector contains sufficient information to simulate the priming effect obtained by Till, Mross and Kintsch (1988, Experiment 1). In that experiment, the final word of each sentence prime was a homograph, and responses were made to target words considered to be related to the distinct meanings of the homograph. Landauer and Dumais demonstrated that the meaning of the context was captured by the centroid vector representation by measuring the similarity between the centroid vector and the vectors for words standing for the different meanings of the homograph

In the studies summarized above, the relationship *between* high-dimensional lexical representations (i.e. vector similarity) was the quantity of interest. What is the nature of the information contained in a *single* word vector? Besides specifying a word's location in high-dimensional semantic space, a co-occurrence vector encodes the frequency distribution of the words occurring in the immediate context of the word of interest. The properties of this distribution vary between words in interesting ways. We demonstrate below that distributional statistics constitute a unique and useful source of information about word use. Furthermore, we shall show that an information-theoretic measure of the characteristics of a word's contexts of use – a measure we term *Contextual Distinctiveness* (CD) – is a psychologically relevant and objective predictor of lexical processing behavior.

In this paper we describe in detail the methodology behind the computation of CD, and in Experiment 1, we investigate its psychological relevance.

CONTEXTUAL DISTINCTIVENESS

Below, we demonstrate how words can be distinguished according to their distributional properties. We first describe between-word differences in subjective

terms, and then seek to objectively measure these differences using the tools of information theory.

A co-occurrence vector summarizes a word's distributional profile – which words it tends to occur with, and how often. For instance, if word *w* tends to appear in a wide variety of linguistic contexts (e.g. *run*), we would expect the distribution of the words it occurs with to be rather diffuse. Conversely, if *w* typically appears in a small number of different contexts (e.g. *amok*), or is perhaps found in diverse contexts but is much more common in a subset of them, its contextual distribution would be less diffuse and we could describe *w* as *contextually distinctive*. Seen from another perspective, encountering the word *run* in isolation is not particularly informative about the linguistic contexts it occurs in (since *run* can appear in wide range of contexts), whereas observing *amok* almost certainly brings to mind the verbal context *run*. Because words appear to vary according to informativeness about their contexts of use, this property, contextual distinctiveness, can be treated as a continuous variable.

We use the relative entropy measure from information theory to quantify the subjective notion of contextual distinctiveness (CD). Since CD intuitively refers to the amount of *information* provided by word *w* about its contexts of use, it is operationalized as the relative entropy (or Kullback-Leibler distance) between the distribution of context words occurring in a window of words around *w* (the *posterior* distribution), and the distribution of context words expected when *w* is not taken into account (the *prior* distribution). CD can be understood as the quantity of information conveyed about *w*'s contextual behavior.[1]

Note that the posterior distribution is simply *w*'s co-occurrence vector representation, with counts converted to conditional probabilities, and the prior is the distribution of context words based on their estimated independent probabilities of occurrence (or relative frequencies) in a large corpus. The prior can be interpreted as the probability distribution expected if the corpus had been created from word tokens randomly chosen according to their relative frequencies in a real corpus. Such a corpus of randomly chosen tokens would be completely unstructured, unlike natural language, and consequently co-occurrence vectors extracted from this corpus would simply record the relative frequencies of the context words (i.e. no linguistic dependencies between words would be encoded).

The formal definition of CD is as follows. First, for the prior distribution, we define *C* to be a discrete random variable ranging over the alphabet of symbols $\{c_1, \ldots, c_n\}$, with probability mass function $P(c)$. (The alphabet is the finite set of context words that label the components of a co-occurrence vector.) The values of $P(c_i)$ are obtained using the maximum likelihood estimator:

$$P(c_i) = \frac{f(c_i)}{\sum_{j=1}^{n} f(c_j)}$$

In this equation, $f(c_i)$ is the frequency of c_i in the corpus, and the denominator is the summed corpus frequency of the *n* words in the alphabet. Note that the prior distribution is a true probability mass function, since $P(c_1)+P(c_2)\ldots+P(c_n)=1$.

We next describe the posterior distribution (the distribution of context words given that target word *w* occurs) using the probability mass function $P(c|w)$. Each conditional probability $P(c_i|w)$ is derived from the co-occurrence frequency of the target word *w*

with the context word c_i, normalized by the total co-occurrences of w with each possible symbol c:

$$P(c_i|w) = \frac{f(c_i, w)}{\sum_{j=1}^{n} f(c_j, w)}$$

Finally, CD is calculated as the relative entropy between the two probability mass functions $P(c)$ and $P(c|w)$, using the convention 0 log 0 = 0 (justified by continuity):

$$D\big(P(c) \,||\, P(c|w)\big) = \sum_{i=1}^{n} P(c_i|w) \log_2 \frac{P(c_i|w)}{P(c_i)}$$

Formally, CD measures the quantity of information provided about a random variable (the contexts that word w appears in) by an event (observing word w). Since the above equation uses base 2 logarithms, the units of information are expressed in *bits*.[2]

Creating co-occurrence vectors

Construction of a semantic space model requires a large amount of natural language output; the choice of this corpus has an impact on the psychological plausibility of the resulting model. We used the 10.3 million word spoken language part of the British National Corpus (BNC; Burnage & Dunlop, 1992). The spoken subcorpus (henceforth *BNC-spoken*) consists of a mixture of speech genres sampled from demographic and context-governed sources, including transcripts of unscripted informal conversation, radio programs and government meetings.

We chose to use spoken language as the source of distributional statistics for two reasons. First, spoken language forms the primary environment for human language learning, and for adult language use for most of the population. Children's exposure to speech compared with written text is crucially much larger. Although vocabulary size undoubtedly increases more rapidly through reading, the core vocabulary items and their contexts of use are acquired through the vast amount of speech that children hear. Words have much more opportunity to be learned through spoken language than through written sources, and their semantic representations would be expected to reflect spoken language context to a greater extent. Second, because of the smaller type:token ratio for spoken language, a single word type is encountered, on average, an order of magnitude more often than a single word type in written language. This means that spoken language, in general, provides a more reliable source of contextual information for a given word, which is advantageous both for acquiring word meaning and for constructing reliable co-occurrence representations.

To prepare the BNC-spoken for the extraction of co-occurrence statistics, we took the following steps. First, the corpus was filtered to remove punctuation and SGML markup, retaining only the words together with their part of speech tags. Next, uppercase and lowercase type were conflated to lowercase. Finally, the corpus was lemmatized by mapping each word form to its corresponding *lexeme* in the CELEX lexical database (Baayen, Piepenbrock & Van Rijn, 1993), and then replacing the word in the corpus with its lexeme's canonical form. This was done by mapping the tagset used to assign a part-of-speech label to each word in the BNC, to the much smaller set of part-of-speech categories employed by CELEX

As a result of lemmatizing the corpus, the counts for all inflectional variants of a word are collapsed together into a single lexeme count. For example, *walk*, *walking*, *walked* and *walks* all share their high-dimensional vector representation, labelled with their canonical form *walk*; and similarly there is only one vector component label <walk> corresponding to all four variants. Other morphological variants conflated by lemmatisation are noun plural suffixes (e.g. *cats*), and comparative and superlative adjective forms (e.g. *cleaner*, *cleanest*). Lemmatisation was motivated by the observation that meaning is normally preserved across the inflectional variants of a lexeme, whereas derivational morphological variants are often separated by semantic drift (e.g. *steer*, *steerage*). We assume that a co-occurrence vector – serving as a representation of word meaning – should contain information that is relevant to a more abstract level of representation than the surface word form. The psycholinguistic literature contains long-standing examples (chiefly involving visual word recognition) of processing seemingly being partially determined by the stems of morphologically complex words (e.g. Taft, 1979).

Creating lexical representations from co-occurrence statistics requires decisions to be made about a number of model parameters. The CD measure requires estimation of two probability distributions; consequently, CD values will vary depending on how model parameters are set. As a result, the success of CD as a predictor of processing behavior will also vary. The window size and the number of context words used to define the representational space are two of the most important parameters (e.g. Patel, Bullinaria & Levy, 1998); we set these parameters empirically, by recording the amount of visual lexical decision response time variance explained by CD, for a sample of words tested in our laboratory (see McDonald, 2000). Based on the results of this procedure, we set the window size to five words before and five words after the word of interest, and recorded co-occurrences with the 500 most frequent content words[3] in the BNC-spoken in order to define the posterior distribution component of the CD measure.

It would be useful to verify that CD really does capture the subjective concept of contextual distinctiveness described earlier. The spoken language component of the BNC contains disfluencies such as filled pauses (*ah*, *erm*, *hmm*, etc) – these would be expected to be among the least contextually distinctive tokens in the corpus, since they can occur in virtually any linguistic environment. This expectation was verified using the objective CD measure; *er* and *erm* received the first and third lowest CD scores of the entire lexicon (0.041 and 0.046 bits, respectively).

In order to graphically illustrate the distributional differences that underlie different CD values for words matched on other lexical properties, we selected two words of equivalent lexeme frequency and plotted their prior and posterior probability distributions (Figures 1 and 2). *Lane* and *customer* are both unambiguously nouns, according to CELEX, and have BNC-spoken lexeme frequencies of 613 and 614, respectively. However, *lane* is substantially higher in CD than *customer* (1.027 bits compared with 0.524 bits). Note that the posterior distribution of *lane* diverges from the prior to a greater degree than does the posterior of *customer*, reflecting, in part, the fact that *lane* occurs in a number of common collocations, such as *back lane* and *fast lane*. CD measures a dimension of lexical variation not captured by simple frequency of occurrence; even though two words may be matched on corpus frequency, the frequency distribution of their co-occurring words can be very different.

RELIABILITY OF CD

A co-occurrence vector records the distributional profile of a word in a corpus, and is therefore merely an estimate of its 'true' distribution in unlimited natural language. Since calculation of CD depends crucially on the distributional information contained in a sample of language, it is necessary to first assess its reliability. Reliability addresses the question of replicability using a different source of data: assuming all else to be equal, would one obtain nearly identical CD scores for a particular word from two different corpora?

CD values computed for common words will be more reliable than for rare words for the same reason that corpus frequency is more reliable for common words: the larger the sample, the more accurately the population value can be estimated (the smaller the measurement error). Two separate estimates of CD for the same high frequency word will be close to the population value, and will therefore be very similar. Because of the unreliability of statistics based on small samples, the results of psycholinguistic studies that use rare words as stimuli are suspect, particularly those using sets of words matched on corpus frequency (Lovelace, 1988). For instance, Gernsbacher (1984) demonstrated that low-frequency words varied substantially in experiential (subjective) familiarity, leading her to reinterpret the results of several word recognition experiments.

It is essential to determine a practical lower limit on the frequency of words for which CD can be confidently measured. Reliability of the CD measure could be estimated by comparing the current CD values obtained using the BNC-spoken to values computed from another corpus of spoken language, for the same set of words. The correspondence between the measurements should be strongest for high-frequency words and fade to nonsignificant levels as frequency drops.

Lacking another comparable speech corpus, we split the BNC-spoken in two and calculated CD for random samples of words taken from a range of frequency intervals. The BNC-spoken was divided into two halves by alternating 10,000 token chunks. Natural log-transformed lexeme frequency in the 10M word BNC-spoken ranged from 0 to 12.963 (see Table 1). We divided this range into 8 equally-spaced intervals, and selected a random sample of 100 words from each bin, with the constraint that the selected words had to appear in both subcorpora; note, though, that bins 1 and 2 consisted of only 21 and 69 words, respectively. Next, we extracted co-occurrence vectors for each word (690 in total) from both subcorpora, using a window size of ± 5, and the 500 most frequent content words in the BNC-spoken as context words.

Next, we calculated CD for each word (690 in total) from each half of the corpus separately, and estimated the reliability of CD for each sample using Kendall's coefficient of concordance *W*. Kendall's *W* will be high when the rank orders of the CD scores derived from each subcorpus are similar. As expected, reliability was very high for the bins containing the most frequent words (see Table 1). The significance of the reliability score for each bin was tested with the X^2 statistic. At the $\alpha=0.01$ level of significance, the null hypothesis that the reliability scores for bins 7 and 8 (the two lowest frequency intervals) were due to chance failed to be rejected.

What did this exercise tell us about the reliability of CD? Clearly, the measure is the most reliable for words for which the most corpus evidence exists. CD could not be reliably measured for words falling in the two lowest bins. Consequently, we do not calculate CD for words with a lexeme frequency of less than 25 occurrences, restricting its applicability to approximately 8,000 lexemes in the BNC-spoken.

EXPERIMENT 1: WORD RECOGNITION

If Contextual Distinctiveness is a psychologically valid lexical property, then between-word differences in CD value should predict differences in processing behavior. In this experiment, we test the hypothesis that the amount of information conveyed by a word about its contextual behavior co-varies with the effort of processing that word: the more information a word carries about its context, the more effort is involved in processing that word. Specifically, we assess the ability of CD to account for subjects' performance on a visual word recognition task. Recognizing the individual words is a necessary first step in reading and understanding a sentence; the process of word recognition arguably involves retrieving the semantic information associated with the word. We have claimed that CD reflects the processing effort involved in recovering word meaning. We therefore predict that the time taken to identify a string of letters as a valid word will vary directly with the quantity of information conveyed by that word about its contexts of use. We re-analysed the data reported by Chumbley and Balota (1984), who employed the standard visual lexical decision (VLD) and naming tasks to produce response times for a large selection of words.

Method

Chumbley and Balota (1984) selected 144 words from the Battig and Montague (1969) norms to serve as stimuli in a series of four word recognition experiments. Of these words, 109 met our frequency-based criterion for calculating reliable CD values (see above). We computed CD for these 109 words, converting seven plural forms to their canonical (singular) forms. A random sample of the stimuli with their corresponding CD values is presented in Table 2.

Mean VLD and naming latencies (for separate groups of 24 subjects) for these words were supplied by Chumbley and Balota (1984, Appendix). Chumbley and Balota conducted two experiments using the lexical decision task – the second (Experiment 3) used nonword foils that more closely matched the critical lexical stimuli in length. We employed the response times obtained in their Experiment 3 in our analysis.

Results and Discussion

CD was positively correlated with mean VLD latency: Pearson r=0.475, p<0.0001. This correlation confirmed our hypothesis that the amount of information a word conveys about its contexts of use is predictive of the effort of processing that word. Words that appear in relatively constrained contexts have high Contextual Distinctiveness values and tend to produce longer lexical decision latencies. In contrast, words whose contexts of use are unconstrained have low CD scores, and less time is required to classify them as real English words.

CD was significantly correlated with naming latency: r=0.431, p<0.0001. The amount of information a word conveys about its contextual behavior is also predictive of the effort involved in pronouncing that word. The recognition of an isolated word has been argued to be influenced by its meaning; CD is able to capture the effect of between-word differences in meaning on word recognition processes using the completely objective source of information available in a record of language output.

CONCLUSIONS

Previous corpus-based studies (e.g. Lund & Burgess, 1996) have demonstrated that the meaning of a word is intimately related to the linguistic contexts in which it is used. A co-occurrence vector representation for a word carries useful information about its meaning, and provides an objective means for estimating the semantic similarities and differences between words. This co-occurrence-based approach to semantic representation constitutes a new and psychologically interesting methodology for investigating word meaning, but application thus far has been restricted to quantifying the relationships involving two or more words. We have probed the nature of the vector representations themselves using the tools of information theory, and have shown that by analysing the contextual distribution encoded in a word vector, it is straightforward to quantify the distributional differences between words. Words vary in the amount of information they convey about their contexts of use – in both the intuitive sense of informativeness and the formal mathematical sense of information. The CD measure summarizes a word's contextual behavior in a single number.

In this paper, we have described in detail the rationale and methodology behind the computation of the CD measure, and have empirically established a frequency-based threshold for application of CD. We have determined that CD measurements are not reliable for lexemes occurring less than 25 times in the 10 million word BNC-spoken, but it is possible to calculate CD for some 8,000 lexemes. This number is still a substantial portion of the lexicon, and is easily sufficient to permit the further experimental investigation of the role of CD in lexical processing.

We have shown that lexical representations constructed from simple co-occurrence statistics contain useful information about a word's contextual behaviour. Contextual Distinctiveness may be readily calculated from large language corpora, and constitutes a new dimension of lexical variation that is relevant to language processing behavior.

REFERENCES

Baayen, R. H., Piepenbrock, R. & van Rijn, H. (1993). *The CELEX lexical database (CD-ROM)*. Philadelphia, PA: Linguistic Data Consortium, University of Pennsylvania.

Battig, W. F. & Montague, W. E. (1969). Category norms for verbal items in 56 categories: A replication and extension of the Connecticut category norms. *Journal of Experimental Psychology Monographs, 80*(3, Pt. 2).

Burnage, G. & Dunlop, D. (1992). Encoding the British National Corpus. In *Papers from the 13th International Conference on English Language Research on Computerised Corpora*.

Cann, R. (*in press*). Functional versus lexical: a cognitive dichotomy. In R. D. Borsley (Ed.) *The nature and function of syntactic categories (Syntax and Semantics 32)*. London: Academic Press.

Chumbley, J. I. & Balota, D. A. (1984). A word's meaning affects the decision in lexical decision. *Memory & Cognition*, *12*, 590-606.

Gernsbacher, M. A. (1984). Resolving 20 years of inconsistent interactions between lexical familiarity and orthography, concreteness, and polysemy. *Journal of Experimental Psychology: General*, *113*, 256-281.

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.

Lovelace, E. A. (1988). On using norms for low-frequency words. *Bulletin of the Psychonomic Society*, *26*, 410-412.

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.

Lund, K., Burgess, C. & Audet, C. (1996). Dissociating semantic and associative word relationships using high-dimensional semantic space. In *Proceedings of the 18th Annual Conference of the Cognitive Science Society* (pp. 603-608). Mahwah, NJ: Erlbaum.

McDonald, S. (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.

Miller, G., Beckwith, R., Fellbaum, C., Gross, D. & Miller, K. (1990). Introduction to WordNet: an online lexical database. *International Journal of Lexicography, 3*, 235-244.

Patel, M., Bullinaria, J. A. & Levy, J. P. (1998). Extracting semantic representations from large text corpora. In J. A. Bullinaria, D. Glasspool & G. Houghton (Eds.)

Proceedings of the 4th Neural Computation and Psychology Workshop, London, 9-11 April 1997 (pp. 199-212). London: Springer-Verlag.

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

Resnik, P. S. (1993). *Selection and information: a class-based approach to lexical relationships*. Unpublished PhD dissertation, University of Pennsylvania.

Spence, D. P., & Owens, K. C. (1990). Lexical co-occurrence and association strength. *Journal of Psycholinguistic Research*, *19*, 317-330.

Taft, M. (1979). Recognition of affixed words and the word frequency effect. *Memory & Cognition*, *7*, 263-272.

Till, R. E., Mross, E. F. & Kintsch, W. (1988). Time course of priming for associate and inference words in discourse context. *Memory & Cognition*, *16*, 283-299.

Table 1. Reliability of CD for Eight Samples.

Bin	Log Frequency Range	1st Word in Bin	N	Sample Size	Kendall $\underline{W}$	X^2
1	12.963-11.344	be	21	21	0.998	39.92*
2	11.343-9.724	yeah	69	69	0.996	135.42*
3	9.723-8.103	work	195	100	0.994	196.73*
4	8.102-6.483	case	810	100	0.971	192.20*
5	6.482-4.862	goal	2204	100	0.939	185.86*
6	4.861-3.242	valid	4624	100	0.789	156.21*
7	3.241-1.621	zebra	9150	100	0.566	112.07
8	1.620-0.000	zulu	28378	100	0.604	119.51

* Significant at α=0.01

Table 2. Contextual Distinctiveness (CD) Values for 10 Lexemes Randomly Selected from Chumbley and Balota (1984).

Lexeme	CD (bits)
brain	0.629
father	0.753
beer	0.908
tea	1.368
chapel	1.483
hunting	1.541
gem	2.368
ruby	2.537
tornado	2.845
tuna	3.007

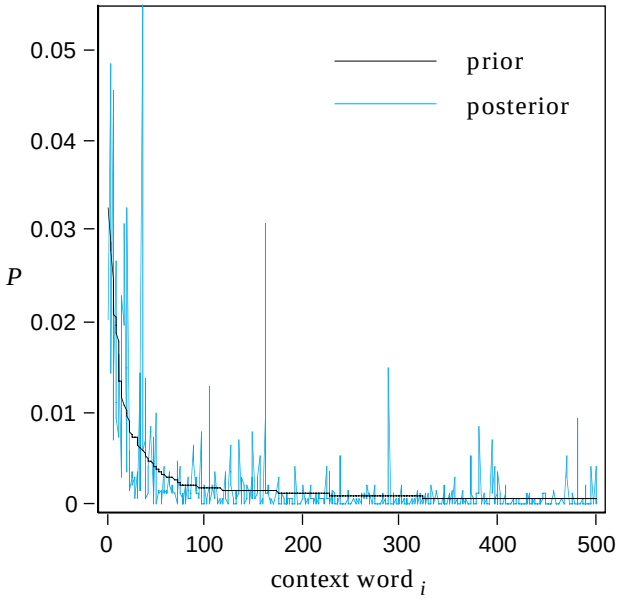

Figure 1. Prior and posterior probability distributions for *lane*. (For clarity, lines are used instead of bars.)

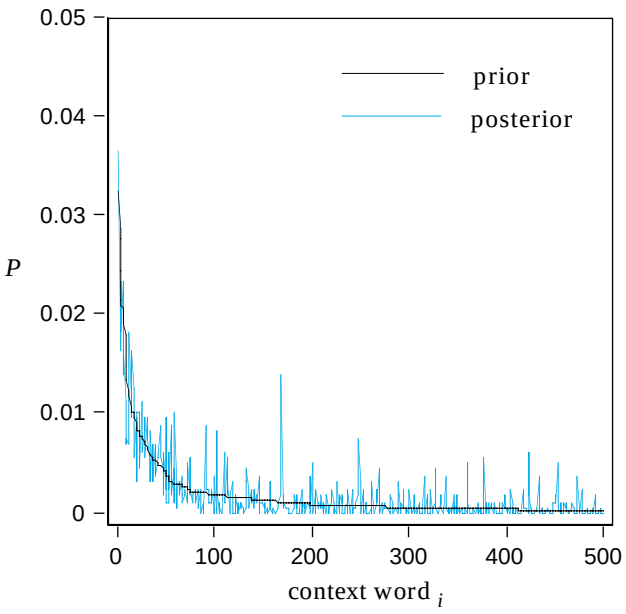

Figure 2. Prior and posterior probability distributions for *customer*.

1 The efficacy of the relative entropy measure to capture distributional differences of this sort has been shown in related work by Resnik (1993), who used relative entropy to estimate the selectional preference strength of a verb for its arguments. However, Resnik defines the prior and posterior distributions in terms of the taxonomically-organised semantic classes in the WordNet lexical database (Miller, Beckwith, Fellbaum, Gross & Miller, 1990), rather than using a finite set of co-occurring words, as is done here.

2 It is important to note that the CD measure is implicitly conditioned on the parameters used when constructing co-occurrence vector representations: window size, the selection of context words, and the choice of corpus. Varying these parameter settings will yield different values of CD for the same word, to a limited extent.

3 We exclude the class of function words from consideration as vector components. There is a growing literature on the processing and representational differences between functional and contentive expressions (Cann, *in press*), which makes distinguishing them in a model of lexico-semantic representation psychologically attractive. Function words can be viewed as the 'building blocks' of syntax, and indeed the use of a context word set consisting primarily of function words to define the axes of the high-dimensional space has led to successful induction of syntactic categories from distributional statistics (e.g. Redington, Chater & Finch, 1998).