Determinants of Adjective-Noun Plausibility

Determinants of Adjective-Noun Plausibility

Maria Lapata and **Scott McDonald** and **Frank Keller**
School of Cognitive Science
Division of Informatics, University of Edinburgh
2 Buccleuch Place, Edinburgh EH8 9LW, UK
{mlap, scottm, keller}@cogsci.ed.ac.uk

Abstract

This paper explores the determinants of adjective-noun plausibility by using correlation analysis to compare judgements elicited from human subjects with five corpus-based variables: co-occurrence frequency of the adjective-noun pair, noun frequency, conditional probability of the noun given the adjective, the log-likelihood ratio, and Resnik's (1993) selectional association measure. The highest correlation is obtained with the co-occurrence frequency, which points to the strongly lexicalist and collocational nature of adjective-noun combinations.

1 Introduction

Research on linguistic plausibility has focused mainly on the effects of argument plausibility during the processing of locally ambiguous sentences. Psycholinguists have investigated whether the plausibility of the direct object affects reading times for sentences like (1). Here, argument plausibility refers to "pragmatic plausibility" or "local semantic fit" (Holmes et al., 1989), and judgements of plausibility are typically obtained by asking subjects to rate sentence fragments containing verb-argument combinations (as an example consider the bracketed parts of the sentences in (1)). Such experiments typically use an ordinal scale for plausibility (e.g., from 1 to 7).

(1) a. [The senior senator **regretted the decision**] had ever been made public.
b. [The senior senator **regretted the reporter**] had ever seen the report.

The majority of research has focussed on investigating the effect of rated plausibility for verb-object combinations in human sentence processing (Garnsey et al., 1997; Pickering and Traxler, 1998). However, plausibility effects have also been observed for adjective-noun combinations in a head-modifier relationship.

Murphy (1990) has shown that *typical* adjective-noun phrases (e.g., *salty olives*) are easier to interpret in comparison to *atypical* ones (e.g., *sweet olives*). Murphy provides a schema-based explanation for this finding by postulating that in typical adjective-noun phrases, the adjective modifies part of the noun's schema and consequently it is understood more quickly, whereas in atypical combinations, the adjective modifies non-schematic aspects of the noun, which leads to interpretation difficulties.

Smadja (1991) argues that the reason people prefer *strong tea* to *powerful tea* and *powerful car* to *strong car* is neither purely syntactic nor purely semantic, but rather lexical.

A similar argument is put forward by Cruse (1986), who observes that the adjective *spotless* collocates well with the noun *kitchen*, relatively worse with the noun *complexion* and not all with the noun *taste*. According to Cruse, words like *spotless* have idiosyncratic collocational restrictions: differences in the degree of acceptability of the adjective and its collocates do not seem to depend on the meaning of the individual words.

1.1 Motivation

Acquiring plausibility ratings for word combinations (e.g., adjective-noun, verb-object, noun-noun) can be useful in particular for language generation. Consider a generator which has to make a choice between *spotless kitchen* and *flawless kitchen*. An empirical model of plausibility could predict that *spotless kitchen* is a plausible lexical choice, while *flawless kitchen* is not.

Adjective-noun combinations can be hard to generate given their collocational status. For a generator which selects words solely on semantic grounds without taking into account lexical constraints, the choice between *spotless kitchen* and *flawless kitchen* may look equivalent. Current work in natural language generation (Knight and Hatzivassiloglou, 1995; Langkilde and Knight, 1998) has shown that corpus-based knowledge can be used to address lexical choice non-compositionally.

In the work reported here we acquire plausibility ratings for adjective-noun combinations by eliciting judgements from human subjects, and examine the extent to which different corpus-based models correlate with human intuitions about the "goodness of fit" for a range of adjective-noun combinations.

The research presented in this paper is similar in motivation to Resnik's (1993) work on selectional restrictions. Resnik evaluated his information-theoretic model of selectional constraints against human plausibility ratings for verb-object combinations, and showed that, in most cases, his model assigned higher selectional association scores to verb-object combinations which were judged more plausible by human subjects.

We test five corpus-based models against human plausibility judgements:

1. **Familiarity of adjective-noun pair.** We operationalise familiarity as co-occurrence frequency in a large corpus. We calculate the co-occurrence frequency of adjective-noun pairs in order to examine whether high corpus frequency is correlated with plausibility, and correspondingly low corpus frequency with implausibility.

2. **Familiarity of head noun.** We compare rated plausibility with the corpus frequency of the head noun, the motivation being that highly frequent nouns are more familiar than less frequent ones, and consequently may affect the judged plausibility of the whole noun phrase.

3. **Conditional probability.** Our inclusion of the conditional probability, $P(noun \mid adjective)$, as a predictor variable also relies on the prediction that plausibility is correlated with corpus frequency. It differs from simple co-occurrence frequency in that it additionally takes the overall adjective frequency into account.

4. **Collocational status.** We employ the log-likelihood ratio as a measure of the collocational status of the adjective-noun pair (Dunning, 1993; Daille, 1996). If we assume that plausibility differences between *strong tea* and *powerful tea* or *guilty verdict* and *guilty cat* reflect differences in collocational status (i.e., appearing together more often than expected by their individual occurrence frequencies), as opposed to being semantic in nature, then the log-likelihood ratio may also predict adjective-noun plausibility.

5. **Selectional association.** Finally, we evaluate plausibility ratings against Resnik's (1993) measure of selectional association. This measure is attractive because it combines statistical and knowledge-based methods. By exploiting a knowledge-based taxonomy, it can capture conceptual information about lexical items and hence can make predictions about word combinations which have not been seen in the corpus.

In the following section we describe our method for eliciting plausibility judgements for adjective-noun combinations. Section 3 reports the results of using the five corpus-based models as predictors of adjective-noun plausibility. Finally, section 4 offers some discussion of future work, and section 5 concluding remarks.

2 Collecting Plausibility Ratings

In order to evaluate the different corpus-based models of adjective-noun plausibility introduced above, we first needed to establish an independent measure of plausibility. The standard approach used in experimental psycholinguistics is to elicit judgements from human subjects; in this section we describe our method for assembling the set of experimental materials and collecting plausibility ratings for these stimuli.

2.1 Method

Materials and Design. The ideal test of any of the proposed models of adjective-noun plausibility will be with randomly-chosen materials. We chose 30 adjectives according to a set of minimal criteria (detailed below), and paired each adjective with a noun selected randomly from three different frequency ranges, which were defined by co-occurrence counts in the 100 million word British National Corpus (BNC; Burnard (1995)). The experimental design thus consisted of one factor, Frequency Band, with three levels (High, Medium, and Low).

We chose the adjectives to be minimally ambiguous: each adjective had exactly two senses according to WordNet (Miller et al., 1990) and was unambiguously tagged as "adjective" 98.6% of the time, measured as the number of different part-of-speech tags assigned to the word in the BNC. The 30 adjectives ranged in BNC frequency from 1.9 to 49.1 per million.

We identified adjective-noun pairs by using Gsearch (Corley et al., 1999), a chart parser which detects syntactic patterns in a tagged corpus by exploiting a user-specified context free grammar and a syntactic query. Gsearch was run on a lemmatised version of the BNC so as to compile a comprehensive corpus count of all nouns occurring in a modifier-head relationship with each of the 30 adjectives. Examples of the syntactic patterns the parser identified are given in Table 1. From the syntactic analysis provided by the parser we extracted a table containing the adjective and the head of the noun phrase following it. In the case of compound nouns, we only included sequences of two

nouns, and considered the rightmost occurring noun as the head.

From the retrieved adjective-noun pairs, we removed all pairs where the noun had a BNC frequency of less than 10 per million, as we wanted to reduce the risk of plausibility ratings being influenced by the presence of a noun unfamiliar to the subjects. Finally, for each adjective we divided the set of pairs into three "bands" (High, Medium, and Low), based on an equal division of the range of log-transformed co-occurrence frequency, and randomly chose one noun from each band. Example stimuli are shown in Table 2. The mean log co-occurrence frequencies were 3.839, 2.066 and .258, for the High, Medium, and Low groups, respectively.

30 filler items were also included, in order to ensure subjects produced a wide range of plausibility ratings. These consisted of 30 adjective-noun combinations that were not found in a modifier-head relation in the BNC, and were also judged highly implausible by the authors.

Procedure. The experimental paradigm was magnitude estimation (ME), a technique standardly used in psychophysics to measure judgements of sensory stimuli (Stevens, 1975), which Bard et al. (1996) and Cowart (1997) have applied to the elicitation of linguistic judgements. The ME procedure requires subjects to estimate the magnitude of physical stimuli by assigning numerical values proportional to the stimulus magnitude they perceive. In contrast to the 5- or 7-point scale conventionally used to measure human intuitions, ME employs an interval scale, and therefore produces data for which parametric inferential statistics are valid.

ME requires subjects to assign numbers to a series of linguistic stimuli in a proportional fashion. Subjects are first exposed to a modulus item, which they assign an arbitrary number. All other stimuli are rated proportional to the modulus. In this way, each subject can establish their own rating scale, thus yielding maximally fine-graded data and avoiding the known problems with the conventional ordinal scales for linguistic data (Bard et al., 1996; Cowart, 1997; Schütze, 1996).

In the present experiment, subjects were presented with adjective-noun pairs and were asked to rate the degree of adjective-noun fit proportional to a modulus item. The experiment was carried out using WebExp, a set of Java-Classes for administering psycholinguistic studies over the Word-Wide Web (Keller et al., 1998). Subjects first saw a set of instructions that explained the ME technique and included some examples, and had to fill in a short questionnaire including basic demographic information. Each subject saw all 120 items used in the experiment (3×30 experimental items and 30 fillers).

Subjects. The experiment was completed by 24 unpaid volunteers, all native speakers of English. Subjects were recruited via postings to local Usenet newsgroups.

2.2 Results and Discussion

As is standard in magnitude estimation studies, statistical tests were done using geometric means to normalise the data (the geometric mean is the mean of the logarithms of the ratings). An analysis of variance (ANOVA) indicated that the Frequency Band effect was significant, in both by-subjects and by-items analyses: $F_1(2, 46) = 79.09$, $p < .001$; $F_2(2, 58) = 19.99$, $p < .001$. The geometric mean of the ratings for adjective-noun combinations in the High band was 2.966, compared to Medium items at 2.660 and Low pairs at 2.271.[1] Post-hoc Tukey tests indicated that the differences between all pairs of conditions were significant at $\alpha = .01$, except for the difference between the High and Medium bands in the by-items analysis, which was significant at $\alpha = .05$. These results are perhaps unsurprising: pairs that are more familiar are rated as more plausible than combinations that are less familiar. In the next section we explore the linear relationship between plausibility and co-occurrence frequency further, using correlation analysis.

3 Corpus-based Modelling

3.1 Method

We correlated rated plausibility (Plaus) with the following five corpus-based variables: (1) log-transformed co-occurrence frequency (CoocF), measured as the number of times the adjective-noun pair occurs in the BNC; (2) log-transformed noun frequency (NounF), measured as the number of times the head noun occurs in the BNC; (3) conditional probability (CondP) of the noun given the adjective estimated as shown in equation (2); (4) collocational status,[2] estimated using the log-likelihood statistic (LLRatio); and (5) Resnik's measure of selectional association (SelAssoc), which measures the semantic fit of a particular semantic class c as an argument to a predicate p_i. The selectional association between class c and predicate p_i is given in equations (3) and (4). More specifically, selectional association represents the contribution of a particular semantic class c to the total quantity of information provided by a predicate about the semantic class of its argument, when measured as the relative entropy between the prior distri-

[1] For comparison, the filler items had a mean rating of .998.

[2] Mutual information, though potentially of interest as a measure of collocational status, was not tested due to its well-known property of overemphasising the significance of rare events (Church and Hanks, 1990).

Pattern	Example
adjective noun	educational material
adjective specifier noun	usual weekly classes
adjective noun noun	environmental health officers

Table 1: Example of noun-adjective patterns

Adjective	Co-occurrence Frequency Band					
	High		Medium		Low	
hungry	animal	1.79	pleasure	1.38	application	0
guilty	verdict	3.91	secret	2.56	cat	0
temporary	job	4.71	post	2.07	cap	.69
naughty	girl	2.94	dog	1.6	lunch	.69

Table 2: Example stimuli (with log co-occurrence frequencies in the BNC)

bution of classes $p(c)$ and the posterior distribution $p(c \mid p_i)$ of the argument classes for a particular predicate p_i.

$$(2)\quad P(noun \mid adjective) = \frac{f(adjective, noun)}{f(adjective)}$$

$$(3)\quad A(p_i, c) = \frac{1}{\eta_i} \cdot P(c \mid p_i) \cdot \log \frac{P(c \mid p_i)}{P(c)}$$

$$(4)\quad \eta_i = \sum_c P(c \mid p_i) \cdot \log \frac{P(c \mid p_i)}{P(c)}$$

In the case of adjective-noun combinations, the selectional association measures the semantic fit of an adjective and each of the semantic classes of the nouns it co-occurs with. We estimated the probabilities $P(c \mid p_i)$ and $P(c)$ similarly to Resnik (1993) by using relative frequencies from the BNC, together with WordNet (Miller et al., 1990) as a source of taxonomic semantic class information. Although the selectional association is a function of the predicate and all semantic classes it potentially selects for, following Resnik's method for verb-object evaluation, we compared human plausibility judgements with the maximum value for the selectional association for each adjective-noun combination.

Table 3 shows the models' predictions for three sample stimuli. The first row contains the geometric mean of the subjects' responses.

3.2 Results

The five corpus-based variables were submitted to a correlation analysis (see Tables 5 and 4). The highest correlation with judged plausibility was obtained with the familiarity of the adjective-noun combination (as operationalised by corpus co-occurrence frequency). Three other variables were also significantly correlated with plausibility ratings: the conditional probability $P(noun \mid adjective)$, the log-likelihood ratio, and Resnik's selectional association measure. We discuss each predictor variable in more detail:

1. **Familiarity of adjective-noun pair.** Log-transformed corpus co-occurrence frequency was significantly correlated with plausibility (Pearson $r = .570$, $n = 90$, $p < .01$). This verifies the Frequency Band effect discovered by the ANOVA, in an analysis which compares the individual co-occurrence frequency for each item with rated plausibility, instead of collapsing 30 pairs together into an equivalence class. Familiarity appears to be a strong determinant of adjective-noun plausibility.

2. **Familiarity of head noun.** Log frequency of the head noun was not significantly correlated with plausibility ($r = .098$), which suggests that adjective-noun plausibility judgements are not influenced by noun familiarity.

3. **Conditional probability.** The probability of the noun given the adjective was significantly correlated with plausibility ($r = .220$, $p < .05$). This is unsurprising, as conditional probability was also correlated with co-occurrence frequency ($r = .497$, $p < .01$).

4. **Collocational status.** The log-likelihood statistic yielded a significant correlation with plausibility ($r = .350$, $p < .01$), a fact that supports the collocational nature of plausible adjective-noun combinations. The log-likelihood ratio was in turn correlated with co-occurrence frequency ($r = .725$, $p < .01$) and conditional probability ($r = .405$, $p < .01$).

5. **Selectional association.** Resnik's measure of selectional association was also significantly correlated with plausibility ($r = -.269$, $p < .05$).

	hungry animal	hungry application	hungry pleasure
Plaus	3.02	1.46	1.31
CoocF	1.79	1.38	0
NounF	9.63	9.69	8.67
CondP	.003	.002	.0005
LLRatio	26.81	14.33	2.9
SelAssoc	.5	.5	.22

Table 3: Models' prediction for *hungry* and its three paired noun heads

However, it should be noted that selectional association was *negatively* correlated with plausibility, although Resnik found the measure was positively correlated with the judged plausibility of verb-object combinations, consistent with its information-theoretic motivation. Resnik's metric was also negatively correlated with co-occurrence frequency ($r = -.226$, $p < .05$), but there was no correlation with noun frequency, conditional probability, or log-likelihood ratio.

Since several of the corpus-based variables were intercorrelated, we also calculated the squared semipartial correlations between plausibility and each corpus-based variable. This allows the unique relationship between each predictor and plausibility (removing the effects of the other independent variables) to be determined. Co-occurrence frequency accounted uniquely for 15.52% of the variance in plausibility ratings, while noun frequency, conditional probability, log-likelihood ratio, and selectional association accounted for .51%, .53%, .41% and 1.7% of the variance, respectively. This confirms co-occurrence frequency as the best predictor of adjective-noun plausibility.

One explanation for the negative correlation between selectional association and plausibility, also pointed out by Resnik, is the difference between verb-object and adjective-noun combinations: combinations of the latter type are more lexical than conceptual in nature and hence cannot be accounted for on purely semantic or syntactic grounds. The abstraction provided by a semantic taxonomy is at odds with the idiosyncratic (i.e., lexical) nature of adjective-noun co-occurrences. Consider for instance the adjective *hungry*. The class ⟨entity⟩ yields the highest selectional association value for the highest rated pair *hungry animal*. But ⟨entity⟩ also yields the highest association for the lowest rated pair *hungry application* (A(*hungry*, ⟨entity⟩) = .50 in both cases). The highest association for *hungry pleasure*, on the other hand, is given by the class ⟨act⟩ (A(*hungry*, ⟨act⟩) = .22). This demonstrates how the method tends to prefer the most frequent classes in the taxonomy (e.g., ⟨entity⟩, ⟨act⟩) over less frequent, but intuitively more plausible classes (e.g., ⟨feeling⟩ for *pleasure* and ⟨use⟩ for *application*).

This is a general problem with the estimation of the probability of a class of a given predicate in Resnik's method, as the probability is assumed to be uniform for all classes of a given noun with which the predicate co-occurs. Although the improvements suggested by Ribas (1994) try to remedy this by taking the different senses of a given word into account and implementing selectional restrictions in the form of weighted disjunctions, the experiments reported here indicate that methods based on taxonomic knowledge have difficulties capturing the idiosyncratic (i.e., lexicalist) nature of adjective-noun combinations.

Finally, idiosyncrasies in WordNet itself influence the performance of Resnik's model. One problem is that sense distinctions in WordNet are often too fine-grained (Palmer (1999) makes a similar observation). Furthermore, there is considerable redundancy in the definition of word senses. Consider the noun *application*: it has 27 classes in WordNet which include ⟨code⟩, ⟨coding system⟩, ⟨software⟩, ⟨communication⟩, ⟨writing⟩ and ⟨written communication⟩. It is difficult to see how ⟨code⟩ or ⟨coding system⟩ is not ⟨software⟩ or ⟨writing⟩ is not ⟨written communication⟩. The fine granularity and the degree of redundancy in the taxonomy bias the estimation of the frequency of a given class. Resnik's model cannot distinguish classes which are genuinely frequent from classes which are infrequent but yet overly specified.

4 Future Work

Although familiarity of the adjective-noun combination proved to be the most predictive measure of judged plausibility, it is obvious that this measure will fail for adjective-noun pairs that never co-occur at all in the training corpus. Is a zero co-occurrence count merely the result of insufficient evidence, or is it a reflection of a linguistic constraint? We plan to conduct another rating experiment, this time with a selection of stimuli that have a co-occurrence frequency of zero in the BNC. These data will allow a further test of Resnik's selectional association measure.

	Plaus	CoocF	NounF	CondP	LLRatio	SelAssoc
Min	.770	0	6.988	.0002	.02	.100
Max	3.240	5.037	11.929	.2139	1734.88	1.000
Mean	2.632	2.054	9.411	.0165	176.24	.288
Std Dev	.529	1.583	1.100	.0312	334.23	.170

Table 4: Descriptive statistics for the six experimental variables

	Plaus	CoocF	NounF	CondP	LLRatio
CoocF	.570**				
NounF	.098	.221*			
CondP	.220*	.497**	.008		
LLRatio	.350**	.725**	.001	.405**	
SelAssoc	−.269*	−.226*	−.191	−.097	.015
*$p < .05$ (2-tailed)			**$p < .01$ (2-tailed)		

Table 5: Correlation matrix for plausibility and the five corpus-based variables

We also plan to investigate the application of similarity-based smoothing (Dagan et al., 1999) to zero co-occurrence counts, as this method is specifically aimed at distinguishing between unobserved events which are likely to occur in language from those that are not. Plausibility ratings provide a suitable test of the psychological validity of co-occurrence frequencies "recreated" with this method.

5 Conclusions

This paper explored the determinants of linguistic plausibility, a concept that is potentially relevant for lexical choice in natural language generation systems. Adjective-noun plausibility served as a test bed for a number of corpus-based models of linguistic plausibility. Plausibility judgements were obtained from human subjects for 90 randomly selected adjective-noun pairs. The ratings revealed a clear effect of familiarity of the adjective-noun pair (operationalised by corpus co-occurrence frequency).

In a correlation analysis we compared judged plausibility with the predictions of five corpus-based variables. The highest correlation was obtained with the co-occurrence frequency of the adjective-noun pair. Conditional probability, the log-likelihood ratio, and Resnik's (1993) selectional association measure were also significantly correlated with plausibility ratings. The correlation with Resnik's measure was negative, contrary to the predictions of his model. This points to a problem with his technique for estimating word class frequencies, which is aggravated by the collocational nature of noun-adjective combinations.

Overall, the results confirm the strongly lexicalist and collocational nature of adjective-noun combinations. This fact could be exploited in a generation system by taking into account corpus co-occurrence counts for adjective-noun pairs (which can be obtained straightforwardly) during lexical choice. Future research has to identify how this approach can be generalised to unseen data.

Acknowledgements

The authors acknowledge the support of the Alexander S. Onassis Foundation (Lapata), the UK Economic and Social Research Council (Keller, Lapata), the Natural Sciences and Engineering Research Council of Canada, and the ORS Awards Scheme (McDonald).

References

Ellen Gurman Bard, Dan Robertson, and Antonella Sorace. 1996. Magnitude estimation of linguistic acceptability. *Language*, 72(1):32–68.

Lou Burnard, 1995. *Users Guide for the British National Corpus*. British National Corpus Consortium, Oxford University Computing Service.

Kenneth Ward Church and Patrick Hanks. 1990. Word association norms, mutual informations, and lexicography. *Computational Linguistics*, 16(1):22–29.

Martin Corley, Steffan Corley, Matthew W. Crocker, Frank Keller, and Shari Trewin, 1999. *Gsearch User Manual*. Human Communication Research Centre, University of Edinburgh.

Wayne Cowart. 1997. *Experimental Syntax: Applying Objective Methods to Sentence Judgments*. Sage Publications, Thousand Oaks, CA.

D. A. Cruse. 1986. *Lexical Semantics*. Cambridge Textbooks in Linguistics. Cambridge University Press, Cambridge.

Ido Dagan, Lillian Lee, and Fernando Pereira. 1999. Similarity-based models of word cooccurrence probabilities. *Machine Learning*, 34(1).

Béatrice Daille. 1996. Study and implementation of combined techniques for automatic extraction of terminology. In Judith Klavans and Philip Resnik, editors, *The Balancing Act: Combining Symbolic and Statistical Approaches to Language*, pages 49–66. MIT Press, Cambridge, MA.

Ted Dunning. 1993. Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, 19(1):61–74.

Susan M. Garnsey, Neal J. Pearlmutter, Elisabeth M. Myers, and Melanie A. Lotocky. 1997. The contributions of verb bias and plausibility to the comprehension of temporarily ambiguous sentences. *Journal of Memory and Language*, 37(1):58–93.

V. M. Holmes, L. Stowe, and L. Cupples. 1989. Lexical expectations in parsing complement-verb sentences. *Journal of Memory and Language*, 28(6):668–689.

Frank Keller, Martin Corley, Steffan Corley, Lars Konieczny, and Amalia Todirascu. 1998. WebExp: A Java toolbox for web-based psychological experiments. Technical Report HCRC/TR-99, Human Communication Research Centre, University of Edinburgh.

Kevin Knight and Vasileios Hatzivassiloglou. 1995. Two-level, many paths generation. In *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics*, pages 252–260, Cambridge, MA.

Irene Langkilde and Kevin Knight. 1998. Generation that exploits corpus-based statistical knowledge. In *Proceedings of the 17th International Conference on Computational Linguistics and 36th Annual Meeting of the Association for Computational Linguistics*, pages 704–710, Montréal.

George A. Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine J. Miller. 1990. Introduction to WordNet: an on-line lexical database. *International Journal of Lexicography*, 3(4):235–244.

Gregory L. Murphy. 1990. Noun phrase interpretation and noun combination. *Journal of Memory and Language*, 29(3):259–288.

Martha Palmer. 1999. Consistent criteria for sense distinctions. *Computers and the Humanities*, to appear.

Martin J. Pickering and Martin J. Traxler. 1998. Plausibility and recovery from garden paths: An eye-tracking study. *Journal of Experimental Psychology: Learning Memory and Cognition*, 24(4):940–961.

Philip Stuart Resnik. 1993. *Selection and Information: A Class-Based Approach to Lexical Relationships*. Ph.D. thesis, University of Pennsylvania.

Francesc Ribas. 1994. On learning more appropriate selectional restrictions. In *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics*, Las Cruces, NM.

Carson T. Schütze. 1996. *The Empirical Base of Linguistics: Grammaticality Judgments and Linguistic Methodology*. University of Chicago Press, Chicago.

Frank Smadja. 1991. Macrocoding the lexicon with co-occurrence knowledge. In Uri Zernik, editor, *Lexical Acquisition: Using Online Resources to Build a Lexicon*, pages 165–189. Erlbaum, Hillsdale, NJ.

Stanley S. Stevens, editor. 1975. *Psychophysics: Introduction to its Perceptual, Neural, and Social Prospects*. John Wiley, New York.