The Institute For Research In Cognitive Science

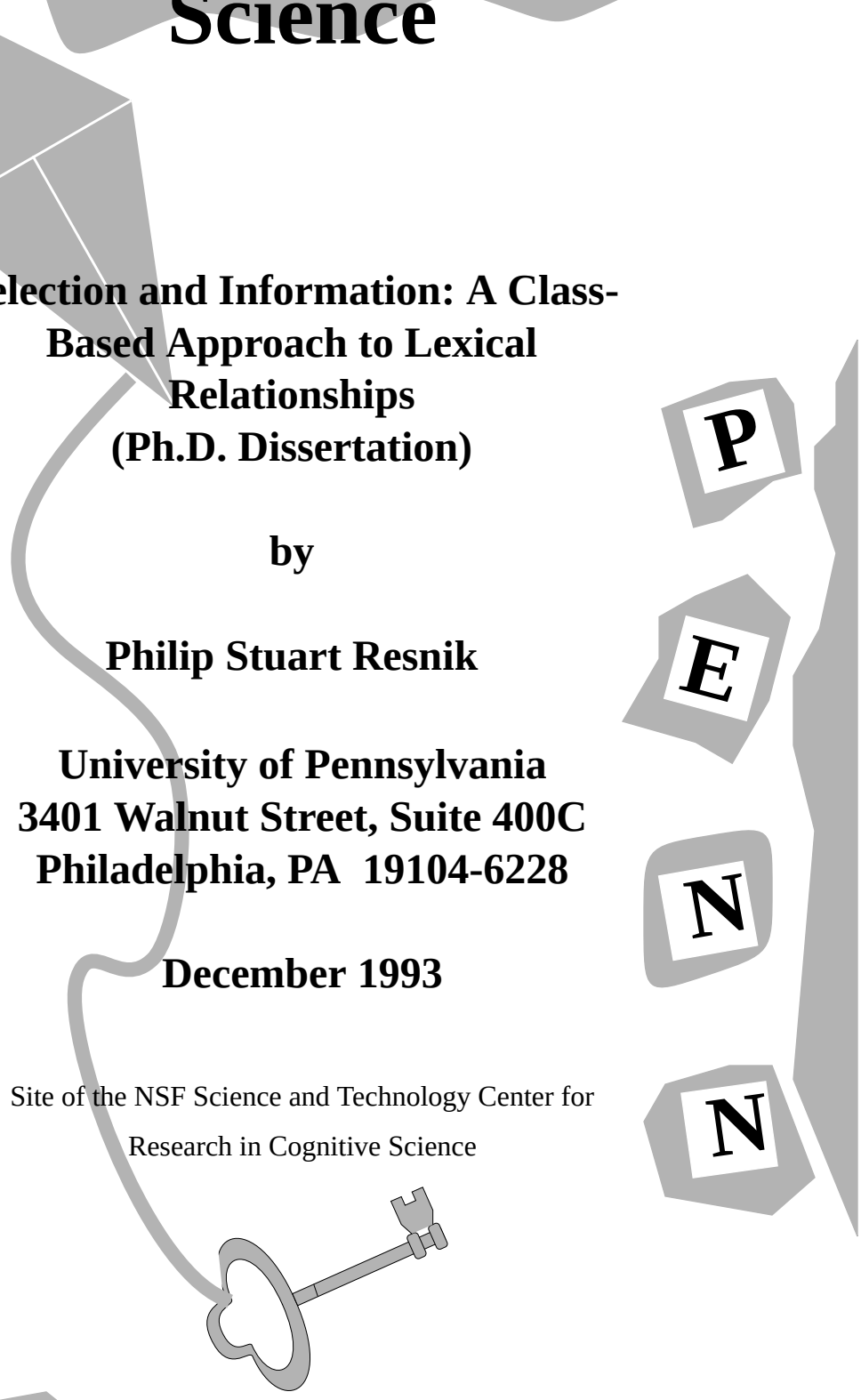
Selection and Information: A Class-Based Approach to Lexical Relationships (Ph.D. Dissertation)

by

Philip Stuart Resnik

University of Pennsylvania
3401 Walnut Street, Suite 400C
Philadelphia, PA 19104-6228

December 1993



Site of the NSF Science and Technology Center for Research in Cognitive Science

University of Pennsylvania
Founded by Benjamin Franklin in 1740

IRCS Report 93-42

SELECTION AND INFORMATION:
A CLASS-BASED APPROACH TO LEXICAL RELATIONSHIPS

Philip Stuart Resnik

A dissertation
in
Computer and Information Science

Presented to the Faculties of the University of Pennsylvania in Partial Fulfillment of the Requirements for the Degree of Doctor of Philosophy

1993

Aravind Joshi
Supervisor of Dissertation

Mark Steedman
Graduate Group Chairperson

For Michael Resnik

Abstract

Selection and Information:
A Class-Based Approach to Lexical Relationships

Philip Stuart Resnik

Supervisor: Aravind Joshi

Selectional constraints are limitations on the applicability of predicates to arguments. For example, the statement "The number two is blue" may be syntactically well formed, but at some level it is anomalous — BLUE is not a predicate that can be applied to numbers.

According to the influential theory of (Katz and Fodor, 1964), a predicate associates a set of defining features with each argument, expressed within a restricted semantic vocabulary. Despite the persistence of this theory, however, there is widespread agreement about its empirical shortcomings (McCawley, 1968; Fodor, 1977). As an alternative, some critics of the Katz-Fodor theory (e.g. (Johnson-Laird, 1983)) have abandoned the treatment of selectional constraints as semantic, instead treating them as indistinguishable from inferences made on the basis of factual knowledge. This provides a better match for the empirical phenomena, but it opens up a different problem: if selectional constraints are the same as inferences in general, then accounting for them will require a much more complete understanding of knowledge representation and inference than we have at present.

The problem, then, is this: how can a theory of selectional constraints be elaborated without first having either an empirically adequate theory of defining features or a comprehensive theory of inference?

In this dissertation, I suggest that an answer to this question lies in the representation of conceptual knowledge. Following Miller (1990b), I adopt a "differential" approach to conceptual representation, in which a conceptual taxonomy is defined in terms of inferential relationships rather than definitional features. Crucially, however, the inferences underlying the stored knowledge are not made explicit. My hypothesis is that a theory of selectional constraints need make reference only to knowledge stored in such a taxonomy, without ever referring overtly to inferential processes. I propose such a theory, formalizing selectional relationships in probabilistic terms: the selectional behavior of a predicate is modeled as its distributional effect on the conceptual classes of its arguments. This is expressed using the information-theoretic measure of relative entropy (Kullback and Leibler, 1951), which leads to an illuminating interpretation of what selectional constraints are: the *strength* of a predicate's selection for an argument is identified with the quantity of *information* it carries about that argument.

In addition to arguing that the model is empirically adequate, I explore its application to two problems. The first concerns a linguistic question: why some transitive verbs permit implicit direct objects ("John ate $\emptyset$") and others do not ("*John brought $\emptyset$"). It has often been observed informally that the omission of objects is connected to the ease with which the object can be inferred. I have made this observation more formal by positing a relationship between selectional constraints and inferability. This predicts (i) that verbs permitting implicit objects select more strongly for (i.e. carry more information about) that argument than verbs that do not, and (ii) that strength of selection is a predictor of how often verbs omit their objects in naturally occurring utterances. Computational experiments confirm these predictions.

Second, I have explored the practical applications of the model in resolving syntactic ambiguity. A number of authors have recently begun investigating the use of corpus-based lexical statistics in automatic

parsing; the results of computational experiments using the present model suggest that many lexical relationships are better viewed in terms of underlying conceptual relationships. Thus the information-theoretic measures proposed here can serve not only as components in a theory of selectional constraints, but also as tools for practical natural language processing.

Acknowledgements

In the past, I have occasionally read the acknowledgements in other people's papers and dissertations and thought, well, they really do seem to have thrown in the kitchen sink, haven't they? Which of those people *really* had something significant to do with this work?

I will never think that thought again. Having sat down to acknowledge my debt to the people around me in making this dissertation happen, I realize that the number of people who have had a real influence is enormous, and that simply listing their names is an expedient but criminally understated way of recognizing their contribution.

I am very grateful to my advisor, Aravind Joshi, for his support, for his guidance, and for his role (together with his co-director, Lila Gleitman) in creating the Institute for Research in Cognitive Science. I'm fortunate to have been a part of IRCS at such an exciting time.

I would like to thank the members of my dissertation committee: Steve Abney, Lila Gleitman, Mark Liberman, and Mitch Marcus. They are individually extraordinary, and together they form a committee with enormous personality and intellect.

I would like to thank the participants in Penn's Computational Linguistics Feedback Forum (CLiFF group) — which is to say my fellow grad students, the IRCS postdocs, and the natural language faculty at Penn — for their support and constructive criticism. In particular, this work has profited from discussions with Eric Brill, Barbara DiEugenio, Bob Frank, Michael Hegarty, Jamie Henderson, Shyam Kapur, Libby Levison, Dave Magerman, Michael Niv, Sandeep Prasada, Owen Rambow, Robert Rubinoff, Giorgio Satta, Jeff Siskind, Mark Steedman, Lyn Walker, Mike White, Bonnie Webber, and David Yarowsky. I have also had extremely helpful discussions with Kevin Atteson, Ken Church, Ido Dagan, Christiane Fellbaum, Jane Grimshaw, Marti Hearst, Donald Hindle, Tony Kroch, Annie Lederer, Robbie Mandelbaum, Gail Mauner, George Miller, Max Mintz, Fernando Pereira, and Stuart Shieber. I know it's a long list, but each name I look at calls to mind an important discussion, a shared insight, or, more often then not, a whole blur of images and conversations over time.

I'm grateful to have been a part of the Gleitmans' "cheese" seminar, with thanks especially to Henry Gleitman, Lila Gleitman, Mike Kelly, and Barbara Landau. Those meetings are, I think, the very essence of what research is about. I hope that someday I can manage to come close to recreating something like it for other generations of students.

I owe a debt of gratitude to Nan Biltz, Carolyn Elken, Dawn Greisbach, Chris Sandy, Estelle Taylor, and Trisha Yannuzzi for all their help with the ins and outs of the department and the university. Ditto for the computational support of Mark-Jason Dominus, Mark Foster, Ira Winston, and Martin Zaidel.

I would like to thank George Miller and the WordNet “lexigang” for their continued interest and for making WordNet freely available.

I would like to acknowledge helpful conversations with my IBM fellowship technical liaison, Wlodek Zadrozny, and to express my gratitude to Peter Brown, Stephen Della Pietra, Vincent Della Pietra, Fred Jelinek, and Bob Mercer for the enormous amount I learned working with them at the IBM T. J. Watson Research Center.

That’s just the debts I’ve incurred during four years at Penn. I also owe a great many thanks to the people who helped me get started in research as an undergraduate — in particular, Barbara Grosz, John Whitman, and Bill Woods. (Thanks, too, to Laurence Bouvard, for helping to inspire my interest in linguistics.) And thanks to those I worked with and learned from at Bolt Beranek and Newman — in particular, Rusty Bobrow, Bob Ingria, Lance Ramshaw, and Ralph Weischedel.

As for the personal debts, words on paper seem especially inadequate. The support and love of my parents are constants that I could not have done without. And it feels to me as if my old friends (especially Debbie Co, Dan Josell, and Lynn Stein) and new friends (especially Howard Lang, Libby Levison, Robbie Mandelbaum, and Owen Rambow) have saved my life more times than I can count.

Finally: Tracy and Benjamin. Words could not possibly say.

Oh yes, and then there’s money. This work was partially supported by the following grants: ARO DAAL 03-89-C-0031, DARPA N00014-90-J-1863, NSF IRI 90-16592, and Ben Franklin 91S.3078C-1, and by an IBM Graduate Fellowship.

Contents

Abstract **iv**

Acknowledgements **vi**

1 Introduction **1**
1.1 Setting . . . 1
1.2 Argument . . . 2
1.3 Chapter Summaries . . . 3

2 Word Classes in Corpus-Based Research **5**
2.1 Overview . . . 5
2.2 Lexical Statistics and their Limitations . . . 6
2.3 Word Classes Based on Lexical Distributions . . . 9
2.3.1 Smoothing methods . . . 10
2.3.2 Proximity methods and clustering . . . 12
2.3.3 Vector representations . . . 14
2.3.4 Discussion . . . 16
2.4 Word Classes Based on a Taxonomy . . . 21
2.4.1 The WordNet noun taxonomy and its semantics . . . 22
2.4.2 Estimation of class probabilities . . . 27
2.4.3 Comparison with distributional methods . . . 29

3 An Information-Theoretic Account of Selectional Constraints **34**
3.1 Overview . . . 34
3.2 Category Mistakes . . . 36
3.2.1 Entailment . . . 36
3.2.2 Presupposition and meaninglessness . . . 37
3.2.3 Implicatures, pragmatics, and metalinguistic negation . . . 38
3.3 Selection Restrictions . . . 42
3.3.1 Selection restrictions as lexical features . . . 42
3.3.2 Selection restrictions as syntactic features . . . 43
3.3.3 Selection restrictions as semantic constraints . . . 44
3.3.4 Selection restrictions and inference . . . 45
3.4 Summary and Prospects . . . 48

3.4.1 Properties of selectional constraints . . . 48
3.4.2 A dilemma . . . 49
3.4.3 A proposal . . . 51
3.5 Selection as Information . . . 53
3.5.1 Intuitions . . . 53
3.5.2 Formalization . . . 54
3.6 Predicted Behavior . . . 59
3.7 Empirical Behavior . . . 61
3.7.1 Computational apparatus . . . 61
3.7.2 Traditional examples . . . 63
3.7.3 Argument plausibility . . . 68
3.8 Other Computational Approaches . . . 72
3.9 Summary . . . 75

4 Selectional Preference and Implicit Objects 76
4.1 Overview . . . 76
4.2 Implicit Object Alternations . . . 77
4.2.1 Non-lexically conditioned object omission . . . 78
4.2.2 Lexically-conditioned object omission . . . 78
4.2.3 Diagnostics . . . 80
4.2.4 Properties of implicit objects . . . 81
4.3 Experiment 1: Selection and Optionality . . . 84
4.3.1 Procedure . . . 84
4.3.2 Results . . . 85
4.3.3 Discussion . . . 86
4.4 Experiment 2: Selection and Frequency of Omission . . . 87
4.4.1 Procedure . . . 88
4.4.2 Results . . . 88
4.5 Experiment 3: Distinguishing Subclasses of Object-drop Verbs . . . 88
4.6 General Discussion . . . 90
4.6.1 Aspectual constraints . . . 91
4.6.2 Taxonomic relationships . . . 94
4.6.3 Summary . . . 97
4.7 Thoughts on verb acquisition . . . 98
4.7.1 Plausibility considerations . . . 98
4.7.2 Relation to bootstrapping . . . 99

5 Semantic Classes and Syntactic Ambiguity 103
5.1 Overview . . . 103
5.2 Parsing Preference Strategies . . . 104
5.3 Coordination . . . 105
5.3.1 Cues to the correct analysis . . . 105
5.3.2 Approximating the cues . . . 106
5.3.3 Experiment 1 . . . 111

5.3.4 Experiment 2 . . . 113
5.4 Prepositional Phrase Attachment . . . 114
5.4.1 Lexical association . . . 115
5.4.2 Prepositional objects . . . 116
5.4.3 Conceptual association . . . 116
5.4.4 Experimental results . . . 120
5.4.5 Relation to other work . . . 123
5.5 Nominal Compounds . . . 126
5.5.1 Syntactic bias and semantic preferences . . . 126
5.5.2 Implementation . . . 127
5.5.3 Quantitative Evaluation . . . 127
5.5.4 Qualitative Evaluation . . . 128

6 Conclusions **132**
6.1 Contributions . . . 132
6.2 Thoughts on Future Work . . . 133

A Notes on Probability Estimation **134**
A.1 Unit Credit Assignment . . . 134
A.2 Good-Turing Estimates . . . 134
A.3 Frequency Estimates Using the Taxonomy . . . 136

B Experimental Data from Chapter 4 **138**
B.1 Experiment 1, Brown Corpus . . . 138
B.2 Experiment 1, CHILDES . . . 139
B.3 Experiment 1, Norms . . . 139
B.4 Experiment 2, Brown Corpus . . . 140
B.5 Experiment 2, CHILDES Corpus . . . 140
B.6 Experiment 2, Norms . . . 141
B.7 Experiment 3, Verbs from Lehrer's (1970) Verb Classification . . . 141
B.8 Experiment 3, Brown Corpus . . . 142
B.9 Experiment 3, CHILDES . . . 143

C Word Similarity Data from Chapter 5 **144**

Bibliography **146**

Name Index **158**

List of Tables

2.1 Verb-object pairs for *drink* (with count > 1) . . . 8
2.2 Verb-object pairs for *open* (with count > 1) . . . 8
2.3 Schematic representation of a lexical matrix . . . 24
2.4 Brown corpus nouns missing from WordNet . . . 27
3.1 Selectional association for NP-bias verbs . . . 70
3.2 Selectional association for clausal-bias verbs . . . 71
4.1 Subclassification of object-drop verbs . . . 89
4.2 Some verbs and their associated classes of direct objects . . . 96
5.1 Suffix rules for reducing nouns to root form. . . . 107
5.2 Superclasses for $\langle nickel, 3567117 \rangle$ and $\langle dime, 3567068 \rangle$ . . . 108
5.3 Similarity with *tobacco* computed by maximizing information . . . 109
5.4 Incorrect bracketings of [n1 [n2 n3]] . . . 129
5.5 Incorrect bracketings of [[n1 n2] n3] . . . 130

List of Figures

2.1 Relationship between mutual information and entropy 7
3.1 Example of a prior distribution and a posterior distribution 55
3.2 Simple distributions to illustrate relative entropy 57
4.1 Selectional behavior of *need* and *eat* 83
4.2 Correlation between selection and implicit objects 89
4.3 Schematic view of the verb mapping problem 99
5.1 Noun classes in prepositional phrase attachment 117
A.1 Example of smoothing in Good-Turing probability estimation 135
A.2 A simple taxonomy 136

Chapter 1

Introduction

1.1 Setting

This thesis is about lexical relationships. Its underlying premise is that the information-theoretic view of language as a stochastic phenomenon and the linguistic view of language as a cognitive phenomenon, though often characterized as being in opposition to each other, are not fundamentally incompatible. Although this premise is accepted in principle by many, it seems only rarely to have found its way into actual research.

I demonstrate the compatibility of the two viewpoints by showing that selectional constraints — long discussed by linguists and philosophers of language — can be expressed as an information-theoretic relationship in a way that respects those discussions rather than ignoring them. I argue that this formalization has value both for linguistic analysis and for practical work in natural language processing (NLP).

In the process of doing this work, a priority of mine has been that the ideas be coherent both to researchers in statistical NLP methods and also to linguists and psycholinguists interested in language from a cognitive perspective. Walking this line has not always been easy, especially with regard to methodology — I expect that some cognitive scientists will remain unconvinced by experiments that use on-line text corpora rather than human subjects, and that some applications-oriented NLP researchers will question the value of building introspective knowledge into a system without regard to what the actual training and test data are going to be. Nonetheless, I hope even those people will find the result relevant and interesting, if not ultimately persuasive.

The proposals in this thesis are, I think, consistent with directions in which research on language from both practical and theoretical perspectives appears to be evolving. I will mention just a few examples of what I mean by this. First, it is becoming clear that statistical methods in natural language processing are moving toward the integration of more linguistic information into probabilistic models — as an indication of how much so, consider that the Penn Treebank is moving in the direction of annotating not only surface linguistic structure but predicate-argument structure, as well (Marcus, Santorini, and Marcinkiewicz, 1993). This makes perfect sense, since the value of a probabilistic model is ultimately constrained by how well its underlying structure — that is, the event space over which it is defined — matches the underlying structure of the phenomenon it is modeling. Despite references to "purely statistical" models of language, there is no such thing: even the simple n-gram model has underlying it a finite-state model, entailing a commitment to the view of linguistic structure criticized to such great effect by Chomsky in *Syntactic Structures*.

Second, in studies of human sentence processing there is an increasing interest in models based on the satisfaction of probabilistic constraints rather than the application of rules or strategies; see, for example, (MacDonald, in press; Tabossi et al., in press). Such studies share many of the same concerns that are central in constructing stochastic language models: how plausible is this lexical combination as compared to that one, and given this context what is expected next? In addition, I think work in language learning is beginning to pay increasing attention to large quantities of realistic data. This is reflected in empirical studies that use the CHILDES data collection to confirm or refute hypotheses (e.g. (Xu and Pinker, 1992)) and in theoretical work on language learning that takes the messy nature of real data into account (e.g. (Siskind, 1993a; Kapur, 1992)).

1.2 Argument

Selectional constraints are limitations on the applicability of predicates to arguments. For example, the statement "The number two is blue" may be syntactically well formed, but at some level it is anomalous — BLUE is not a predicate that can be applied to numbers. Philosophers have called examples like this one "category mistakes," and generative linguists have called them "selectional violations."

The most influential theory of selectional constraints has been the one proposed by Katz and Fodor (1964), according to which a predicate associates a set of defining features with each argument, expressed within a restricted semantic vocabulary. Despite the persistence of this theory, however, there is widespread agreement about its empirical shortcomings (McCawley, 1968; Fodor, 1977). As an alternative, some critics of the Katz-Fodor theory (e.g. (Johnson-Laird, 1983)) have abandoned the treatment of selectional constraints as semantic, instead treating them as indistinguishable from inferences made on the basis of factual knowledge. This provides a better match for the empirical phenomena, but it opens up a different problem: if selectional constraints are the same as inferences in general, then accounting for them will require a much more complete understanding of knowledge representation and inference than we have at present.

The problem, then, is this: how can a theory of selectional constraints be elaborated without first having either an empirically adequate theory of semantic features or a comprehensive theory of conceptual knowledge and inference?

I will suggest that an answer to this question lies in the representation of conceptual knowledge. Following Miller (1990b), I adopt a "differential" approach to conceptual representation, in which a conceptual taxonomy is defined in terms of inferential relationships rather than definitional features. Specifically, knowledge about words is represented in terms of other words whose meanings they share. I characterize the notion of "sharing" a meaning in terms of plausible entailments: two words share a meaning if there is a representative context in which they are mutually substitutable without changing the inferences one would ordinarily be licensed to draw. Crucially, however, the inferences themselves are not made explicit in the knowledge representation. The role of inferences is indirect — they determine what the structure of the taxonomy will be, but otherwise are not a part of the knowledge stored there.

My hypothesis is that a theory of selectional constraints need make reference only to knowledge stored in a taxonomy of this kind, without ever referring overtly to inferential processes or to other forms of factual knowledge. I propose such a theory, formalizing selectional relationships in probabilistic terms. The selectional behavior of a predicate is modeled as its distributional effect on the conceptual classes of its arguments, expressed using the relative entropy between the *prior* distribution of argument concepts and the

posterior distribution of argument concepts given the predicate. Using this information-theoretic measure leads to an illuminating interpretation of what selectional constraints are: the *strength* of a predicate's selection for an argument is identified with the quantity of *information* it carries about that argument.

In the computational implementation of this model, WordNet (Miller, 1990b) serves as a proxy for the conceptual taxonomy, and on-line corpora provide linguistic input. Thus, unlike previous theories, the present model demonstrates the capacity to acquire selectional constraints, and it has been tested using large quantities of naturally occurring data. The performance of the implementation supports the empirical adequacy of the theoretical model: without additional special-purpose algorithms, the implemented model shows appropriate behavior when confronted with traditional examples considered by Katz and Fodor, and their critics.

The remainder of the thesis concerns the application of the theory to two problems. First, I consider a linguistic question: why some transitive verbs permit implicit direct objects ("John ate $\emptyset$") and others do not ("*John brought $\emptyset$"). It has often been observed informally that the omission of objects is connected to the ease with which the object can be inferred. I have made this informal observation more precise by positing a relationship between selectional constraints and inferability. This would predict (i) that verbs permitting implicit objects select more strongly for (i.e. carry more information about) that argument than verbs that do not, and (ii) that strength of selection is a predictor of how often verbs omit their objects in naturally occurring utterances. Computational experiments confirm these predictions.

Second, I explore the practical applications of the model in resolving syntactic ambiguity, following a number of authors (e.g. Hindle and Rooth (1991; 1993)) who have recently begun investigating the use of corpus-based lexical statistics in parsing. The hypothesis considered here is that many lexical relationships reflect underlying conceptual relationships, and that statistical disambiguation strategies should take those into account. Like approaches that create and use word classes on the basis of distributional behavior in text corpora, this provides some measure of resistance to the problem of data sparseness. Unlike those approaches, however, the use of knowledge-based rather than distributional classes provides a clear interpretation for what is in a class, and takes advantage of existing on-line knowledge sources. Although the use of semantic or conceptual word classes in disambiguation has been investigated using a small set of semantic primitives or text from a restricted domain (Basili, Pazienza, and Velardi, 1991; Chang, Luo, and Su, 1992; Grishman and Sterling, 1992; Weischedel et al., 1991), to my knowledge the present work is the first to apply statistical disambiguation techniques using a large-scale conceptual taxonomy to unrestricted text corpora. The results suggest that the information-theoretic measures proposed here can serve not only as components in a theory of selectional constraints, but also as tools for practical natural language processing.

1.3 Chapter Summaries

Chapter 2. In this chapter, I discuss the use of corpus-based statistics to capture lexical properties. After illustrating how some limitations of statistics based solely on word co-occurrence suggest generalizing from words to word classes, I discuss the alternatives of using classes based on distributional similarity and using classes based on taxonomic knowledge. I then propose to compute class-based statistics using a knowledge-based, conceptual taxonomy, detailing the semantics of such a taxonomy and discussing how class-based probabilities are estimated.

Chapter 3. This chapter represents the core of the thesis. Limits on the applicability of predicates to arguments have variously been called sortal constraints, selection (or selectional) restrictions or constraints, and type rules. A review of the literature on such constraints suggests two different ways in which they can be characterized: a "semantic" approach, which has been questioned on empirical grounds, and an "inferential" approach, for which the theoretical issues are at present poorly understood. In this chapter, I propose a new, information-theoretic formalization of selectional constraints based on the taxonomic representation introduced in Chapter 2, and argue that it addresses both theoretical and empirical concerns.

Chapter 4. In this chapter, I investigate one application of the model proposed in Chapter 3, exploring the relationship between selectional constraints and argument omissibility for verbs in English. It has been observed that the ability of some verbs to omit their objects is connected with the inferability of properties for that argument, and that inferability can to a great extent be identified with the selectional information carried by the verb. This hypothesis is supported by a computational study: the first experiment demonstrates that verbs permitting implicit objects tend as a group to select more strongly for that argument than obligatorily transitive verbs; the second experiment demonstrates that the tendency in practice to drop the object of verbs correlates with selectional preference strength; and a third experiment investigates the inferability of direct objects for verbs that do and do not require a salient antecedent for that argument in order for it to be omitted. I conclude the chapter with a discussion of some possible implications of this study for accounts of verb acquisition by children.

Chapter 5. In this chapter, I investigate a second application of the model proposed in Chapter 3, exploring the use of the implemented model as a statistical method for resolving syntactic ambiguity in processing unconstrained text. I argue that a number of "every way ambiguous" constructions — in particular, prepositional phrase attachment, coordination, and nominal compounds — can be resolved by appealing to conceptual relationships such as selectional preference and semantic similarity, and that class-based, information-theoretic formalizations of these notions provide a practical way to do so.

Chapter 6. I summarize the contributions of the dissertation, and present some thoughts on future work.

Chapter 2

Word Classes in Corpus-Based Research

In this chapter, I discuss the use of corpus-based statistics to capture lexical properties. After illustrating how some limitations of statistics based solely on word co-occurrence suggest generalizing from words to word classes, I discuss the alternatives of using classes based on distributional similarity and using classes based on taxonomic knowledge. I then propose to compute class-based statistics using a knowledge-based, conceptual taxonomy, detailing the semantics of such a taxonomy and discussing how class-based probabilities are estimated.

2.1 Overview

It has become common in statistical studies of natural language data to use measures of lexical association to extract useful relationships between words. To take a few examples, (Smadja, 1991) uses lexical association measures to extract collocation information from large corpora for use in language generation, (Church and Hanks, 1989) propose the use of mutual information to estimate word association norms on the basis of lexical co-occurrence, and (Yarowsky, 1993) shows that local word co-occurrences provide reliable cues for sense disambiguation. (See (Church et al., 1991) for a useful overview of statistical techniques for lexical analysis.)

Lexical association has its limits, however, since often either the data are insufficient to provide reliable lexical correspondences, or a task requires more abstraction than solely lexical correspondences permit. In the next section I illustrate these points by looking at one application of lexical association — a proposal by (Hindle, 1990) to use mutual information in capturing predicate-argument relationships. In the sections that follow, I discuss the extension of lexical relationships to class-based relationships, and consider the advantages and disadvantages of constructing word classes on the basis of lexical distributions in corpora. I then turn to the possibility of using word classes defined in terms of a knowledge-based taxonomy. In particular, I consider the theory of lexical representation implemented in WordNet (Beckwith et al., 1991), which is closely related to a proposal by Sparck Jones (1964). The chapter concludes with a straightforward method for estimating probabilities in such a noun-class taxonomy on the basis of lexical co-occurrence in

a corpus; this will lay the groundwork for the information-theoretic model of selectional constraints to be proposed in Chapter 3.

2.2 Lexical Statistics and their Limitations

Recent discussions of lexical statistics often begin with mutual information, an information-theoretic measure of association used with natural language data to gauge the "relatedness" between two words. The mutual information between two words x and y is defined as follows:

$$\mathrm{I}(x;y) \quad = \quad \log \frac{\mathrm{p}(x,y)}{\mathrm{p}(x)\mathrm{p}(y)}. \tag{2.1}$$

Intuitively, the probability of seeing x and y together, $\mathrm{p}(x,y)$, gives some idea as to how related they are. However, if x and y are both very common, then it is likely that they appear together frequently simply by chance and not as a result of any relationship between them. In order to correct for this possibility, $\mathrm{p}(x,y)$ is divided by $\mathrm{p}(x)\mathrm{p}(y)$, which is the probability that x and y would have of appearing together by chance if they were independent. Taking the logarithm of this ratio gives mutual information some desirable properties; for example, its value is respectively positive, zero, or negative according to whether x and y appear together more frequently, as frequently, or less frequently than one would expect if they were independent.

Another quite useful interpretation of mutual information can be derived by looking at the information-theoretic notion of *entropy*. The entropy of a random variable X is defined as the expected value of $-\log \mathrm{p}(x)$. That is,

$$\begin{aligned} \mathrm{H}(X) \quad &= \quad \mathrm{E}\left[-\log \mathrm{p}(x)\right] \\ &= \quad -\sum_{x} \mathrm{p}(x) \log \mathrm{p}(x), \end{aligned} \tag{2.2}$$

where the summation is over all possible values of X. The quantity $\mathrm{H}(X)$ is, roughly speaking, a measure of how uncertain we are about the value that X will have. The *conditional* entropy of X given another random variable Y measures the uncertainty of X, given that the value of Y is known:

$$\begin{aligned} \mathrm{H}(X|Y) \quad &= \quad E\left[-\log \mathrm{p}(x|y)\right] \\ &= \quad -\sum_{x,y} \mathrm{p}(x,y) \log \mathrm{p}(x|y). \end{aligned} \tag{2.3}$$

Clearly, $\mathrm{H}(X|Y)$, the uncertainty about X given that you know the value of Y, is always less than or equal to $\mathrm{H}(X)$, since any additional knowledge about Y can only decrease (or at worst have no effect on) our uncertainty about X.

Now, the mutual information of random variables X and Y is:

$$\begin{aligned} \mathrm{I}(X;Y) \quad &= \quad E\left[\log \frac{\mathrm{p}(x,y)}{\mathrm{p}(x)\mathrm{p}(y)}\right] \\ &= \quad \mathrm{H}(X) - \mathrm{H}(X|Y). \end{aligned} \tag{2.4}$$

Notice that this is just the expected value for the quantity defined in equation (2.1), which is also known more precisely as "pointwise mutual information."

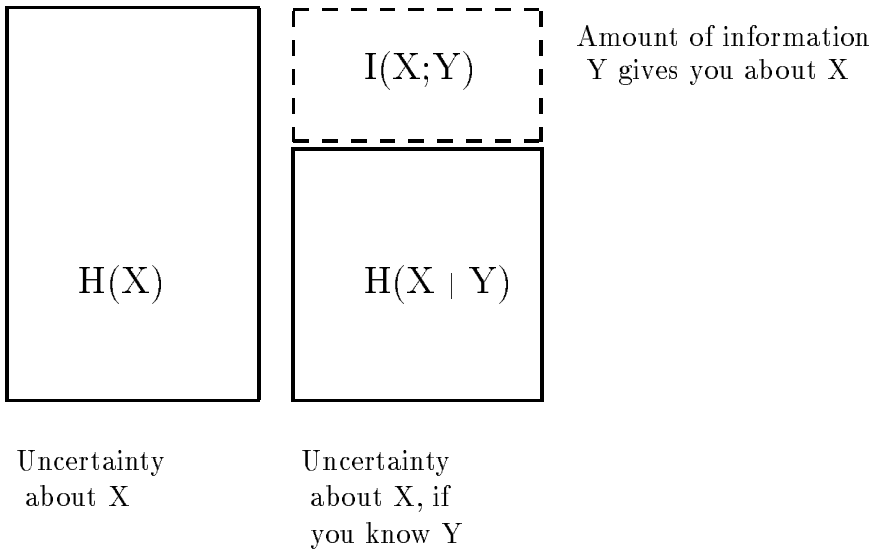


Figure 2.1: Relationship between mutual information and entropy

As equation (2.4) shows, mutual information is a measure of how much information Y provides about X — that is, how much it decreases uncertainty.[1] This relationship between mutual information and entropy is depicted in Figure 2.1.

As an example of how mutual information has been used in corpus-based work, consider Hindle's (1990) application of mutual information to the discovery of predicate-argument relations. Unlike some researchers who restrict themselves to surface distributions of words, Hindle investigates word co-occurrences as mediated by syntactic structure — that is, words x and y are counted as appearing together whenever they stand in a certain syntactic relationship to each other (e.g. subject, object) within a sentence. A six-million-word sample of Associated Press news stories was parsed in order to construct a collection of subject-verb-object instances. On the basis of these data, Hindle calculated a *co-occurrence score* (an estimate of mutual information) for verb-object pairs and verb-subject pairs. Table 2.1 shows the verb-object pairs for the verb *drink* that occurred more than once, ranked by co-occurrence score, "in effect giving the answer to the question 'what can you drink?"' (Hindle, 1990, p. 270).

Hindle's proposal illustrates two limitations of using mutual information between words as a measure of predicate-argument association:

- Sparseness of data: the corpus may fail to provide sufficient information about relevant word-word relationships
- Lack of abstraction: word-word relationships, even those supported by the data, may not be the appropriate relationships to look at for some tasks.

Indeed, any of a family of statistics based solely on lexical co-occurrences — including mutual information, t-score (Church et al., 1990), χ^2, and so forth — suffers from these limitations. Let us consider them in turn.

[1] Being symmetrical, it also measures how much information X provides about Y. See (Cover and Thomas, 1991) for a clear discussion of mutual information and related topics.

score	verb	object
12.34	drink	bunch (of) beer
11.75	drink	tea
11.75	drink	Pepsi
11.75	drink	champagne
10.53	drink	liquid
10.20	drink	beer
9.34	drink	wine
7.65	drink	water
5.15	drink	anything
2.54	drink	much
1.25	drink	it
1.22	drink	SOME AMOUNT

Table 2.1: Verb-object pairs for *drink* (with count > 1)

score	verb	object	count
7.76	open	closet	2
6.93	open	mouth	9
6.79	open	door	32
6.14	open	window	5
5.88	open	store	2
5.76	open	season	2
4.54	open	church	2
4.49	open	heart	2
4.24	open	eye	7
2.38	open	way	4

Table 2.2: Verb-object pairs for *open* (with count > 1)

First, as in all statistical applications, it must be possible to estimate probabilities accurately. Although larger and larger corpora are increasingly available, the specific task under consideration often can restrict the choice of corpus to one that provides a smaller sample than necessary to discover all the lexical relationships of interest. This can lead some lexical relationships to go unnoticed.

For example, the Brown Corpus of American English (Francis and Kučera, 1982) has the attractive (and for some tasks, necessary) property of providing a sample that is balanced across many genres. In an experiment using the Brown Corpus, modeled after Hindle's (1990) investigation of predicate-argument relationships, I calculated the mutual information between verbs and the nouns that appeared as their objects.[2] Table 2.2 shows objects of the verb *open*. As in Table 2.1, the listing includes only verb-object pairs that were encountered more than once.

Attention to the verb-object pairs that occurred *only* once, however, led to an interesting observation. Included among the "discarded" object nouns was the following set: *discourse*, *engagement*, *reply*, *program*, and *session*. Although each of these appeared as the object of *open* only once — too infrequently to provide reliable probability estimates — this set, considered as a whole, reveals an interesting fact about some kinds of things that can be opened, roughly captured by the notion of *communications*. More generally, several

[2] Direct objects in this experiment were identified using the parsed version of the Brown corpus found in the Penn Treebank.

pieces of statistically unreliable information at the lexical level may nonetheless capture a useful statistical regularity when combined. This observation motivates an approach to lexical association that makes such combinations possible.

The second limitation of word-word associations is simply this: some tasks are not amenable to treatment using lexical relationships alone. An example is the automatic discovery of verb argument preferences for natural language systems. Here, the relationship of interest holds not between a verb and a noun, but between the verb and a *class* of nouns (e.g. between *eat* and nouns representing things that are edible). Given a table built using lexical statistics, such as Table 2.1 or 2.2, no single lexical item necessarily stands out as the "preferred" object of the verb — the selectional restriction on the object of *drink* would typically be something like "beverages" or "liquids," not *tea*. Once again, the limitation seems to be one that can be addressed by considering sets or classes of nouns, rather than individual lexical items.

2.3 Word Classes Based on Lexical Distributions

The limitations discussed in the previous section suggest a shift from looking at words to looking at groups or classes of words. A class of words will have attributed to it the accumulated properties of its members, even if observations of the individual members are sparse; in addition, word classes can be used to capture higher level abstractions such as syntactic or semantic features.

A great deal of recent work addresses the creation and use of word classes using lexical distributions in text corpora. The premise behind this approach is that the relatedness of words is reflected by similarities in their distributional contexts, as observed in large collections of naturally occurring text. Church *et al.* (1990, p. 159), discussing statistical methods and linguistic performance, sum up this idea as follows:

> Our approach has much in common with a position that was popular in the 1950s. It was common practice to classify words not only on the basis of their meanings but also on the basis of their co-occurrence with other words. Running through the whole Firthian tradition, for example, is the theme that "You shall know a word by the company it keeps" [Firth, 1957]. Harris's "distributional hypothesis" dates from about the same period. He hypothesized that "the meaning of entities, and the meaning of grammatical relations among them, is related to the restriction of combinations of these entities relative to other entities" ([Harris, 1968], p. 12).

In this section, I will review a number of computational proposals for deriving word classes on the basis of distributional behavior in corpora. These can be broken up according to the following rough classification:

- Smoothing methods
  These methods make implicit use of word classes, but do not represent them explicitly.

- Proximity methods and clustering
  Proximity methods define measures of word similarity. Clustering methods use proximity relationships to form explicit word classes, but do not decompose word tokens into representations that are related to class membership.

- Vector representations
  These methods form decompositional word representations, but may or may not use these representations to derive explicit word classes.

Dividing the distributional techniques in this way is somewhat artificial, since many methods are closely related. For example, given a vector representation, it is always possible to derive a proximity measure such as Euclidean distance. Similarly, given a measure of word proximity, it is always possible to perform a cluster analysis to derive explicit classes, and given a hierarchical clustering, it is always possible to define a similarity measure based on some notion of proximity in the hierarchy. Nonetheless, this classification provides a way to make some helpful distinctions among the current approaches reviewed in the subsections that follow.

2.3.1 Smoothing methods

Smoothing is a general term for the combination of multiple sources of information, often under circumstances in which using the information at a single data point might be undesirable or misleading. For example, given a set of data points plotted in two dimensions, it may be uninformative to simply draw a curve that goes through each individual data point (x, y); such a curve might be jagged and hard to interpret, as opposed to a smooth curve that misses some points but reveals a general trend in the data. In the absence of some parametric model (e.g. finding the parabola that best fits the data), a smooth curve can be obtained by "averaging" each data point together with the nearby points.[3] Thus in a sense a point together with its nearby points constitutes a class, and the properties of the class are derived from the properties of its individual members.

For purposes of language processing, the "points" of interest are typically words or sequences of words. For example, the language model in a speech recognition system generates hypotheses about what the next word will be, on the basis of a previous word sequence. Hypotheses take the form of a conditional probability distribution $\mathrm{p}(W_k|w_1, \ldots, w_{k-1})$, where random variable W_k ranges over possible words to be predicted, and word sequence $w_1, \ldots, w_{k-1}$ is the prior context. In many speech systems, it is assumed that words are produced according to an underlying Markov model, so that the distribution is well approximated by

$$\mathrm{p}(W_k|w_1, \ldots, w_{k-1}) \quad \approx \quad \mathrm{p}(W_k|w_{k-n}, \ldots, w_{k-1}), \tag{2.5}$$

where n is the order of the Markov model. Notice that this approximation implicitly represents a grouping of word sequences into classes, since all sequences for which $w_{k-n}, \ldots, w_{k-1}$ is the same are treated identically. For trigram models, where $n = 2$, this means that the sequences (*I'll,see,him,after*) and (*She,met,him,after*) are equivalent — in both cases *dinner* and *the* are each reasonable hypothesis about the word that will follow. However, often even this equivalence classification of prior word sequences is insufficient, because the data are too sparse to estimate probabilities accurately: even for a corpus of reasonable size, many trigrams do not occur at all.

One method proposed for solving this problem is the technique of interpolated estimation (Jelinek and Mercer, 1980; Bahl, Jelinek, and Mercer, 1983), in which several probability estimators are smoothed together. The central idea behind interpolated estimation is to take a linear combination of several estimates, weighting each according to how reliable it is. That is,

$$\widetilde{\mathrm{p}}(w|\sigma) \quad = \quad \sum_i \lambda_i(\varphi_i(\sigma))\mathrm{p}(w|\varphi_i(\sigma)), \tag{2.6}$$

where w is a possible word to be predicted, σ is the prior context, each φ_i is a different equivalence class function, and $\lambda_i(\varphi_i(\sigma))$ must always sum to 1. As an example, consider the case where $\sigma = w_1, \ldots, w_{k-1}$,

[3] See discussion and references in (Press et al., 1988).

and

$$\begin{aligned}\varphi_1(\sigma) &= w_{k-2}, w_{k-1} \\ \varphi_2(\sigma) &= w_{k-1}.\end{aligned}$$

Here, the first estimator uses two previous words of context, and the second estimator uses just one word of context. Clearly it is better to use more information when it is available, so $\lambda_{\varphi_1(\sigma)}$ will be high relative to $\lambda_{\varphi_2(\sigma)}$ when $\varphi_1(\sigma)$ occurs frequently; when it does not, $\lambda_{\varphi_2(\sigma)}$ will be accordingly higher. Jelinek *et al.* discuss methods for estimating the λ_i using held-out data in order to achieve this behavior.

(Bahl et al., 1989) discuss a related approach to combining estimators, in which the equivalence classes are determined not simply by applying an nth-order Markov assumption for varying n, but by classifying prior contexts using a decision tree. The process of classifying a context sequence $w_1, \ldots, w_{k-1}$ involves starting at the top node of the tree and descending along a path until a leaf is reached. At each node, the choice of which branch to descend is made on the basis of a question at that node (e.g., *is w_{k-1} a determiner?*). Thus the top node of the tree can be viewed as implicitly representing an exhaustive equivalence class, and the nodes along a path represent progressively narrower equivalence classes, comprising those context sequences for which all the questions up to this point have been answered the same way.

There are many ways to construct decision tree classifiers, the most common of which is to start with the full set of training data at the root, and to proceed top down, recursively partitioning nodes according to the partition that optimizes some measure such as information gain (Quinlan, 1990). In (Bahl et al., 1989) the best partition — and hence the choice of equivalence classes created at this branch — is chosen by minimizing the average conditional entropy $\mathrm{H}(W_k|c_i)$ over the equivalence classes $\{c_i\}$ in the partition. The conditional entropy at a leaf measures the uncertainty of the prediction made at that leaf, based on the training data.

Because the training data are partitioned at each branching point, by the time a leaf is reached the class it represents may be too small to support statistically reliable predictions. Accordingly, many decision tree construction algorithms prune paths below a certain point (e.g., (Breiman et al., 1984)), increasing the size of the classes represented by the leaves of the tree. Bahl *et al.* take a different approach: they apply interpolated estimation, using estimators based on the equivalence class represented at each node. Specifically, $\mathrm{p}(w|\sigma)$ is calculated by classifying σ in the tree, and then computing

$$\widetilde{\mathrm{p}}(w|\sigma) = \lambda_r \mathrm{p}(w|\varphi_r(\sigma)) + \cdots + \lambda_l \mathrm{p}(w|\varphi_l(\sigma)), \tag{2.7}$$

where $n_r, \ldots, n_l$ are the nodes on the path that was taken from root to leaf and $\varphi_r, \ldots, \varphi_l$ are the equivalence classes associated with those nodes. As before, the λ_i are estimated on the basis of held-out data, with the end result being that the nodes in a path contribute according to how reliably the prediction at that node can be estimated.

In summary, both n-gram and decision tree word prediction methods make implicit use of equivalence classes of word sequences; in the case of decision-tree techniques, a constructed tree implicitly represents a set of equivalence classes derived according to lexical distributions in the training data. Interpolated estimation is one way to combine predictions made on the basis of different equivalence classes.

(Grishman and Sterling, 1993), investigating the automated acquisition of selectional constraints, apply *co-occurrence smoothing* (Essen and Steinbiss, 1992), a technique in which prediction information for distinct words is combined on the basis of their distributional similarity. A matrix P_C of confusion probabilities is constructed on the basis of lexical distributions: $P_C(w_j|w_i)$ indicates the probability on average that word

w_j will occur in a context in which w_i occurs. Using this matrix, the likelihood of a word w given some context can be estimated robustly by combining the likelihoods of all other words w' given that same context, weighting the contribution of each w' by its confusability with w. For example, a smoothed trigram model would be computed as follows:

$$\widetilde{\mathrm{p}}(w_n|w_{n-2},w_{n-1}) \quad = \quad \sum_{w'_n} P_C(w'_n|w_n)\mathrm{p}(w'_n|w_{n-2},w_{n-1}). \tag{2.8}$$

In general, any property of interest can be smoothed in this fashion: if f is some value based on w, then

$$\tilde{f}(w) \quad = \quad \sum_{w'} P_C(w'|w)f(w'). \tag{2.9}$$

Grishman and Sterling smooth frequencies of relational triples (e.g. [*like,subject,Mary*]) and demonstrate that the smoothed frequencies yield improved performance over unsmoothed frequencies in constraining the output of a robust parser.

Co-occurrence smoothing and interpolated estimation are similar in the sense that each computes a smoothed value by weighting the contributions of multiple estimators. However, in the applications of interpolated estimation described above, each estimator is associated with a set of words (or word sequences) that are treated as equivalent. In contrast, co-occurrence smoothing utilizes not equivalence classes, but a continuous measure of relatedness among the words for which estimated values are combined; in this respect, it strongly resembles proximity methods.

2.3.2 Proximity methods and clustering

Since a number of techniques for measuring the similarity (or dissimilarity) of words can be viewed as representing proximity in some semantic space, I will group such techniques under the label of "proximity methods." Generally, these methods work by using the frequency with which two words appear in the same contexts, together with their frequencies in other contexts, to compute a single value representing their proximity. Proximity need not be symmetric — for example, in (Grishman and Sterling, 1993) the confusion probability $P_C(w_j|w_i)$ need not be the same as $P_C(w_i|w_j)$. In addition to its intuitive plausibility, the idea that shared contexts indicate semantic similarity appears to have some psychological validity: (Miller and Charles, 1991) show that the discriminability of word contexts correlates inversely with semantic similarity ratings.[4]

Given a measure of proximity, clustering techniques are used to organize data into groupings of similar entities. It is important to note that traditional applications of data clustering are used for exploratory analysis of existing data. Jain and Dubes (1988, p. 1) write:

> The objective of cluster analysis is simply to find a convenient and valid organization of the data, not to establish rules for separating future data into categories. Clustering algorithms are geared toward finding structure in the data.

This must be distinguished from the automatic learning of classifiers based on training data. Weiss and Kulikowski (1991, p. 17) write:

[4]Miller and Charles determined contextual similarity by having subjects sort sentential contexts (sentences with the target word blanked out) for pairs of target words. In one experiment the sentences came from the Brown corpus, and, in a replication, a separate group of subjects generated sentences containing the target words. They discuss the comparative advantages and disadvantages of contextual similarity measures based on substitutability versus overlap of co-occurring words.

> The objective of learning classifications from sample data is to classify and predict successfully on new data.

Despite the apparent opposition, of course, there are important relationships between cluster analysis and statistical pattern classification. For example, k-nearest neighbor classifiers in effect make use of proximity to form clusters: a new pattern is classified by first seeking an exact match in the training data, and then, if one is not found, choosing the class that appears most frequently among the k training items that are closest (or most similar) to the new pattern (Weiss and Kulikowski, 1991, p. 70). Most of the work to be described here is intended to be used by first fixing a training corpus, and then applying the resulting word classes to new data, so in general these techniques seem closer to classifier learning than to traditional exploratory data analysis. However, experimental results are often evaluated by inspecting the cohesiveness of the clusters that result from training.

One useful way to organize a discussion of word proximity methods is according to the way they determine that two words occur in similar contexts. Perhaps the simplest interpretation of context is string co-occurrence: one word is said to appear in the context of another if the two words are adjacent, or, more generally, if the context word appears within some fixed distance to the left or right. (Church and Hanks, 1989) observe that this form of context may help lexicographers identify useful semantic classes: they suggest that by ranking the words occurring to the right of a word (in their example, the verb *save*) by mutual information, useful patterns may emerge. In effect, their suggestion amounts to computing a word proximity measure, using the word *save* some distance to the left as the shared context, although no explicit measure of similarity is ever computed.

(Brown et al., 1992) make more direct use of mutual information in determining word classes: they create a hierarchical clustering of words in a vocabulary by first assigning each word to its own cluster, and then merging clusters bottom up, at each point choosing to merge the pairs of clusters for which the loss of average mutual information is least. Here mutual information between clusters is calculated using string adjacency. That is, two clusters c_1 and c_2 co-occur when a pair $w_1 w_2$ occurs in the training text, where $w_1 \in c_1$ and $w_2 \in c_2$ — thus the model underlying the clustering criterion is a bigram model.

Brill (1990; 1991) has also investigated bigram-based word clustering; however, his methods operate by specifying a parameterized measure of word proximity based on similarity of bigram distributions, rather than maximizing mutual information. By varying the strictness of the similarity criterion from more to less stringent, a hierarchical clustering results.

Along similar lines, (Bensch and Savitch, 1992) use string adjacency to determine word co-occurrence, representing the context of each word instance as the pair comprising the preceding and following word. However, unlike most distributional approaches to word clustering, they ignore frequencies of co-occurrence, and compute a score, the Tanimoto coefficient, based simply on the number of shared and unshared contexts. Thus what is relevant is not how often two words occur in a shared context, but whether that context is shared at all. Given a word proximity measure based on the Tanimoto coefficient, Bensch and Savitch create a clustering by first creating a fully connected graph of the words in the vocabulary, with proximities on the arcs, and then constructing the minimum spanning tree for the graph. (In this respect the technique is a specialization of the Pathfinder algorithm (Schvaneveldt, Durso, and Dearholt, 1989), which also begins with a fully connected graph and produces a network structure. Pathfinder permits two user-set parameters to vary, and at one extreme value of the parameters the "minimal" network is produced, consisting of the unique minimal spanning tree if there is one, and the union of all edges in any minimal spanning tree, otherwise.)

String co-occurrence within a window is a very coarse-grained way to capture linguistic relationships, and as the size of the window increases, the amount of noise — that is, the frequency of irrelevant co-occurrences — increases. Given a syntactic analysis, however, it is possible to define word contexts using relationships that are more constrained and linguistically well motivated. For example, (Hindle, 1990) proposes a method for classifying nouns on the basis of the verbs for which they appear as arguments, as determined using a robust parser. The similarity of two nouns n_1 and n_2 with respect to a single verb is defined using mutual information: the "object similarity" (similarity relative to the direct object position) is taken to be zero if one noun has positive mutual information and the other negative mutual information with the verb; otherwise this verb can be considered a shared context, and the strength of similarity in this context is quantified by minimizing the magnitude of the two mutual information values. "Subject similarity" is defined analogously, and overall similarity is defined as the sum of subject and object similarity for the two nouns, summed up over all verbs. Grefenstette (1992) has taken a similar approach, though using a broader set of syntactic relationships, and using the Jaccard measure rather than mutual information. Both Hindle and Grefenstette produce a similarity-based ranking of words for a given noun, though an explicit clustering is never constructed.

(McKeown and Hatzivassiloglou, 1993) apply a distributional method not unlike Grefenstette's to the clustering of adjectives on the basis of the nouns they modify. For purposes of measuring similarity, they employ Kendall's τ coefficient, a non-parametric statistic. Unlike the other distributionally-based clustering methods described here, McKeown and Hatzivassiloglou provide their algorithm with *negative* evidence derived from the corpus using linguistic knowledge: essentially, they assume that two adjectives modifying the same instance of a noun cannot be modifiers on the same scale and therefore should not be grouped together. (For example, the phrase *the tall, dark man* provides evidence that *tall* and *dark* belong in different classes.)

(Pereira, Tishby, and Lee, 1993) also use argument relationships to determine similarity, producing a clustering of nouns based on the verbs for which they appear as direct objects. Words and clusters are represented using the probability distribution for co-occurrence with verbs, and similarity (really dissimilarity) is measured using the relative entropy between distributions. The use of relative entropy yields a useful information-theoretic interpretation of the relationship between a word and a cluster: it measures how costly it would be, in bits of information, to use the distribution associated with the cluster rather than the distribution associated with the noun itself. It is important to note that this method, unlike many of the others, produces clusters that are not discrete: cluster membership is a matter of degree.

2.3.3 Vector representations

In most of the techniques described above, the set of contexts in which a word appears can be thought of as a representation for that word in a "semantic" space, the dimensions of which are identified with the set of words in the vocabulary. For instance, if there are k words in the vocabulary, then the frequency distribution of words appearing to immediately to the left of a word can be represented as a k-ary vector of integers. In (Bensch and Savitch, 1992), where co-occurrence is relevant but frequency of co-occurrence is not, a word's context can be interpreted as a vector of binary features.

Under the heading of "vector representation" methods, I will describe a set of techniques that construct word representations on the basis of lexical distributions. In most cases, the resulting representations arise by means of a reduction in the dimensionality of the semantic space determined by the full vocabulary.

To begin with an exception, however, the mutual information clustering method of Brown *et al.*, described above, can be seen as producing not only a hierarchy of discrete clusters, but also a bit-vector representation for each word in the vocabulary. The derivation of representations from the cluster hierarchy is straightforward: if M is the height of the hierarchy — that is, the length of the longest path from the top of the hierarchy to a word at the bottom — then one need only label each connecting branch in the hierarchy with either a 0 or a 1, according to the direction of the branch. The resulting bit-vector representation of each word is the sequence of bits encountered on the path from the top of the hierarchy to that word, where any sequence of length less than M is padded on the right with sufficiently many 0 bits to make its length exactly M. Notice that the "meaning" of each bit in the representation will depend on the sequence of bits preceding it, so the vector does not really identify a point in an M-dimensional space. Nonetheless, such a representation is useful in contexts where the order-dependence of the bits can be exploited — for example, in a binary decision tree, where classification is accomplished by asking questions one bit at a time.

In the connectionist community, a fair amount of work has taken place under the rubric of "distributed representations" — that is, representing a word (or anything else of interest, for that matter) as a pattern of activity across a set of nodes in a neural network. In many cases, these representations are "subsymbolic," in the sense that no individual node constitutes the representation for a single concept or semantic feature. A thorough description of the connectionist literature on this topic is beyond the scope of this discussion; however, see (Smolensky, 1988) on the "sub-symbolic paradigm," and (Smolensky, Legendre, and Miyata, 1992) for an illuminating discussion on the relationship between connectionism and the study of language, as well as pointers to the connectionist literature.

One interesting and influential example of connectionist work on lexical representation is a study by Elman (1990) on the automatic discovery of syntactic/semantic features for words. Elman proposes an extension of the ubiquitous "feed-forward" network architecture — comprising a layer of input units, one or more layers of hidden units, and a layer of output units — in which the input at any time step includes the activations of the hidden units from the previous time step.[5] This "recurrent" network model is trained on (artificially constructed) sentences, in which each word is represented as a bit vector with a single bit active. Its task is to predict the next word at each time step, on the basis of the information at the input units. Since these encode previous context, training the network is not unlike automatically constructing a kind of Markov model. What is of interest here, however, is not performance on the prediction task, but rather the generalizations the network has been forced to make in order to learn to make its predictions. Elman writes:

> [The] network seems to have learned to approximate the likelihood ratios of potential successors. How has this been accomplished? The input representations give no information (such as form-class) that could be used for prediction. The word vectors are orthogonal to each other. Whatever generalizations are true of classes of words must be learned from the co-occurrence statistics, and the composition of those classes must itself be learned. . . . [One] might expect to see these patterns emerge in the internal representations which the network develops in the course of learning the task. These internal representations are captured by the pattern of hidden unit activations which are evoked in response to each word and its context.

Elman studies these internal representations by constructing for each word a single vector representation, computed by averaging the vectors of hidden-unit activity for the word in each context in which it appears. These vector representations are subjected to a hierarchical cluster analysis. The resulting hierarchy shows

[5] This architecture is a variation on a proposal by Jordan (1986). A similar architecture to Elman's is proposed in (Allen, 1990).

that the similarity structure of representations accords with the syntactic and semantic character of the words they represent — for example, the major distinction is between verbs and nouns, within the verb cluster there are sub-clusters corresponding to different subcategorizations, and within the noun cluster there are divisions and subdivisions according to animacy, humanness, and so forth.

It is important to note that Elman's model appears to depend to a great extent on the orthogonality of the input representations, since if these representations do not start out as equally dissimilar, spurious representational similarities between unrelated words could dominate discovered similarities based on distributional evidence. (In a small study investigating this problem, I found that representing words by non-orthogonal bit-vectors did, in fact, lead to this problem of spurious similarity (Resnik, 1991). Initial experiments indicated that the problem could be overcome by making the hierarchical clustering sensitive not to the distance between internal representations, but to the *relative movement* of word representations in representational space over the course of training.) It is also important to note that Elman's method requires multiple co-occurrences in the training data, and perhaps depends on these to a greater extent than statistical methods where low counts can nonetheless be used to compute reasonable probability estimates. It is this concern, in fact, that motivated Bensch and Savitch to explore a frequency-independent method for word classification.

(Schütze, to appear) has investigated the acquisition of distributed representations on a far larger scale than Elman. The heart of his method is a reduction of dimensionality in a semantic space using singular-value decomposition. Schütze first constructs a large matrix containing frequencies of lexical co-occurrence in a string; a singular-value decomposition is then performed on this matrix, and 97 singular values extracted, resulting in vector representations that constitute points in a 97-dimensional "semantic space."[6] Since the singular-value decomposition of a matrix provides the best possible least-squares approximation to the original matrix using the reduced number of dimensions, the resulting 97-dimension vectors will be similar if the original 5000-dimension vectors were.[7]

2.3.4 Discussion

There has clearly been a great deal of interesting research on the automated acquisition of word classes based on lexical distributions — the discussion above illustrates some real variety among approaches, and it is not exhaustive. The results thus far are promising: in most of the works just described, the authors present automatically derived word classes that on inspection seem entirely plausible.

However, despite their obvious potential, it is difficult to assess the value of these distributional class methods at present, and even when there are clear successes, it is particularly difficult to compare the advantages and disadvantages of alternative proposals. For example, Brown *et al.* (1992) succeed in using classes to reduce the space required for storing a language model, but are less successful in their ultimate goal of increasing the model's predictive power on unseen data. Most of the other methods discussed pursue other goals, with ensuing difficulties in drawing comparisons; for example, although Grishman and Sterling (1993) achieve an improvement in performance according to their evaluation metric — essentially success at identifying unlikely relational triples — those results are not evaluated against any other statistical smoothing method or, indeed, any of the other techniques for making use of lexical class information. Among the other authors presenting formal evaluations of their proposals, we find still other methods for assessment: Pereira

[6] For reasons of practicality co-occurrences were taken with respect to four-letter subsequences rather than entire words in some of Schütze's work; however, his most recent work makes direct use of word co-occurrences, e.g. (Schütze, 1993).

[7] See (Deerwester et al., 1990) for a related application of singular-value decomposition in information retrieval

et al. (1993) measure the relative entropy of held-out test data against the trained cluster models, as well as performance on an artificial verb-choice task, and Schütze (to appear) applies his statistically-derived lexical representations to the task of resolving word sense ambiguity.

This difficulty in evaluating alternative methods is by no means isolated to the problem of deriving and using word classes, and undoubtedly more systematic comparisons will be undertaken as further progress is made. However, in addition to the question of evaluation, several other issues are worth noting.

Computational expense. The derivation of word classes based on lexical distributions can be extremely expensive in computational terms. Brown *et al.* (1992) note that they have no practical method for finding an optimal solution to partitioning words into classes, and that even the suboptimal greedy algorithm they use takes $O(V^5)$ time (where V is the size of the vocabulary) if implemented straightforwardly, and $O(V^3)$ if implemented cleverly. For vocabularies that are still too large to handle, they present a further approximation according to which the assignment of words to classes proceeds incrementally, holding the total number of classes constant as new words are added. Computations of this kind can take anywhere from days to weeks of real time, and runs of several months are not unheard of.[8]

Although there is in principle nothing wrong with algorithms that require long-term computation — especially given rapid increases in computational power — increased computing power may not succeed in catching up with the time complexity of statistical algorithms. More important, it may be that the computationally intensive nature of statistical approaches slows down the rate of progress. Since training a statistical model requires a global computation relative to the entire corpus, and since statistics based on a subset of the corpus may not adequately reflect the whole, testing each new idea requires computation on a large scale. On the other hand, it may be that the automated discovery of information by statistical means is worth the wait, when compared to the even longer time-course of knowledge acquisition by hand. For example, (Magerman, 1993) reports parsing results obtained by statistical methods, trained in weeks or months, that are comparable to the performance of a grammar developed for the same domain by a grammarian over the course of a decade.

Class identifiers. A potential problem with distributionally derived word classes is the fact that, in general, the classes that emerge are not identified with symbolic labels of any kind. If classes are viewed as a means for reducing data sparseness, as in the language model of (Brown et al., 1992) or the selectional constraints of (Grishman and Sterling, 1993), this is not a problem — similarity measures or classifications are simply a means for improving the probability estimates for a given word. When a symbolic component is also involved, though, some method is needed for relating clusters to other information. For example, (Schütze, to appear) manually labels clusters of word representations in order to identify word senses, and in (Schütze, 1993) he takes a similar approach for the identification of part-of-speech tags. The choice of classes to label is made on the basis of a relatively small test set, and it is not clear how such an approach would be applied in a more general setting.

The problem is compounded in cases where the classes themselves are not discrete. Pereira *et al.* (1993), describing some results of their hierarchical clustering, present the nouns that are closest to the centroid of

[8] A quote from (Dagan, 1990) seems fairly representative of the prevailing attitude toward computational time among researchers pursuing statistical approaches: "Although the construction of the full size database [of co-occurrence statistics] is not feasible for us, it is clearly feasible for a large scale project. This is shown by a similar database that was implemented as part of the language model of the IBM speech recognition system. . . . Parsing time depends upon the specific parser that is used, but with current technologies it is reasonable to parse a fairly large corpus *within several months*" (emphasis added).

each cluster — for instance, one cluster of direct objects for the verb *fire* is described by the set {*missile, rocket, bullet, gun*}. However, *every* noun in the vocabulary is a member of this cluster to some degree. If, as Pereira *et al.* suggest, these clusters are to be used in automatically constructing grammars, some method for drawing sharp rather than fuzzy class boundaries seems necessary, and ultimately some way of identifying classes symbolically as components in a grammar will be needed.

The semantics of word classes. Although the representations or classes discovered using distributional methods are often described as "semantic," the information captured by means of statistical distributions often defies simple description. For example, clusters hand-selected by Brown *et al.* as particularly "interesting" include the following (they list just the ten most frequent items):

- feet miles pounds degrees inches barrels tons acres meters bytes . . .
- asking telling wondering instructing informing kidding reminding bothering thanking deposing . . .

These groups are encouragingly coherent and even "semantic" in some sense — but notice that other information is encoded, as well, such as number (plural units of measurement) and inflection (verbs of communication in the progressive).

Among their randomly selected rather than hand-picked classes, it is not clear exactly what information is being captured even in the more coherent cases:

- rise focus depend rely concentrate dwell capitalize embark intrude typewriting . . .
- aware unaware unsure cognizant apprised mindful partakers . . .

For example, with the exception of *rise* and *typewriting*, the clustering of *focus*, *depend*, and so forth seems primarily to capture a set of verbs that tend to be followed by *on*. The distributional property of being followed by *of* seems highly relevant for the rather more related group containing *aware*, *unaware*, and so forth, but also appears to pull in the incongruous *partakers*. Many of the other randomly selected classes appear to have some connecting link among the most frequent items, but become increasingly opaque beyond the first few:

- cost expense risk profitability deferral earmarks capstone cardinality mintage reseller . . .
- force ethic stoppage force's conditioner stoppages conditioners waybill forwarder Atonabee . . .
- industry producers makers fishery Arabia growers addition medalist inhalation addict . . .

Many of the examples selected at random by Schütze (to appear) to illustrate his results have a similar character:

- disable: deter intercept repel halting surveillance shield maneuvers
- kid: dad kidding mom ok buddies Mom Oh Hey hey mama

It would seem that the information captured using these techniques is not precisely syntactic, nor purely semantic — in some sense the only word that appears to fit is *distributional*. Of course, Brown *et al.* do not make any claims to the contrary for the above data. They do, however, propose another technique specifically for the purpose of identifying "semantically sticky" groups of words. Again, the hand-selected set of classes they present are encouraging; for example, the following:

- we our ourselves ours
- question questions asking answer answers answering
- write writes writing written wrote pen
- school classroom teaching grade math
- attorney counsel trial court judge

But again it is not clear what links the members of these classes — *pen* and *writing* are undoubtedly associated in some sense, as are *judge* and *trial* or *math* and *classroom*, but it is difficult to go beyond that to a discussion of what general properties hold of classes discovered by this procedure. The best description seems to be "words that tend to appear in similar contexts," which is no more than a restatement of the method by which the classes were derived.

To be fair, it is far from clear exactly what the criteria of success *should* be for an automatic discovery procedure aimed at identifying "semantic" classes. (Miller, 1971) describes four different general methods used by psychologists to investigate semantic similarity among lexical items, and the psychologists' criteria range from frequency of association to substitutability to co-occurrence. The last of these is essentially the line taken by Brown *et al.* in the semantically "sticky" examples just presented.

Conflation of word senses. Schütze (to appear), describing the results of his experiment on vector representations, comments that "vector representations of words that can be used in a wide variety of contexts are not interesting." He illustrates by showing that the nearest "semantic" neighbors of the verb *keeping*, according to his method, do not form a coherent class on any obvious interpretation of semantic relatedness.

Now, *keeping* is not a semantically vacuous word, and its distribution is far from arbitrary. It is likely that *keeping* appears in similar contexts to *putting* more often than it does to, say, *speaking*:

(1) a. keeping/putting/*speaking them together
    b. keeping/putting/*speaking his hands on his head

On the other hand, unlike *putting*, it is also likely to appear in similar contexts to *retaining*:

(2) a. keeping/retaining/*putting possession of the football
    b. keeping/retaining/*putting a permanent record of the transaction

The real issue seems to be not that the word appears in a wide variety of contexts, but that distributional analysis is being done with respect to the word *token* and not the different *senses* associated with that token.

The same concern arises in many other studies of automatic word classification based on lexical distributions. When Brown *et al.* present a semantic class containing both *school* and *grade*, as described above, the grouping is perceived as semantically coherent because the reader assigns an appropriate interpretation to each word on the basis of the other words. However, it is hard to see how their discrete classification could succeed in grouping those two terms, and at the same time also manage to encode the relationship between *slope* and *grade*, or the relationship between *school* and *hospital*, without inducing spurious similarities. If each token is associated with a single point in semantic space, then words having multiple senses will occupy a point determined by the relative frequencies of the individual senses. Although in many cases multiple word senses share relevant properties — for example the *newspaper* and *term paper* senses of *paper* — in

other instances the single point in semantic space represents an amalgam of properties that may not preserve the relationships associated with component word senses.

The problem is reduced to some extent in (Pereira, Tishby, and Lee, 1993), where a noun can be a member of any number of clusters, each of which to some extent encodes a different sense. Thus the appearance of the noun *rocket* in a cluster with *gun* and *weapon* (things that fire projectiles) does not preclude it from also appearing in a cluster with *shot* and *bullet* (the projectiles themselves). Nonetheless, there still appears to be some "leakage" across the word senses of a single word token. The "distance" between noun n and cluster c is calculated by using relative entropy to compare the verb distribution given n and the verb distribution given c:[9]

$$\begin{aligned} d(n,c) &= \mathrm{D}(p_n \parallel p_c) \\ &= \sum_v \mathrm{p}(v|n) \log \frac{\mathrm{p}(v|c)}{\mathrm{p}(v|n)}. \end{aligned} \tag{2.10}$$

Here each term, measuring the contribution of each verb to the total divergence between the two distributions, is calculated using the probability of the verb given the noun *token*, regardless of the sense in which the noun was being used. This means that frequently used senses of the noun will influence cluster membership, potentially overwhelming the distributional characteristics of less frequent senses.

To be specific, consider a noun that has two senses s_1 and s_2, for which sense s_2 is much more frequent than sense s_1. Suppose that cluster c is, on intuitive grounds, closely related to s_1 but not s_2, and that verb v tends to co-occur with s_1 but not with s_2. For example, n might be *back*, in its senses as a football player (s_1) and a body part (s_2), cluster c might represent the portion of the semantic space shared by appropriate senses of *quarterback*, *kicker*, *receiver*, and so forth, and verb v might be *tackle*. Even if $\mathrm{p}(v|s_1)$ is high, the contribution corresponding to *tackle* in equation (2.10) will be low, since $\mathrm{p}(v|n) \approx \mathrm{p}(v|s_2)$; conversely, the contribution corresponding to the verb *arch* may be high. As a result, it is likely that n will be judged an unlikely member of c.

Now, a more informed model might resolve this problem by including the relationship between nouns and noun senses as part of the distance calculation; for example, calculating class membership by replacing equation (2.10) with

$$d(n,c) = \sum_v \sum_i \mathrm{p}(s_i|n,v)\mathrm{p}(v|s_i) \log \frac{\mathrm{p}(v|c)}{\mathrm{p}(v|s_i)}. \tag{2.11}$$

Such a solution is unfortunately circular for Pereira *et al.*, since identifying the senses $\{s_i\}$ for a given noun is part of their task.

A rather more radical solution is suggested by Elman: abandoning the idea of context-independent word tokens altogether. He comments:

> Conventional wisdom holds that as words are heard, listeners retrieve lexical representations. Although these representations may indicate the contexts in which the words acceptably occur, the representations are themselves context-free. They exist in some canonical form which is consistent across all utterances. . . .
>
> A different image is suggested in the approach taken here. As words are processed there is no separate stage of lexical retrieval. The representations of words (the internal states following input of a word) always reflect the input taken together with the prior state. In this scenario,

[9] The relative entropy alone does not measure extent of class membership: probability of membership is a function of this value.

> words are not building blocks as much as they are cues which guide the network through different grammatical states. Words are distinct from each other by virtue of having different causal properties. (Elman, 1989, p. 23)

On this story — I think of it as "radical polysemy" — it is not words that occupy points in semantic space, nor even word senses, but word instances taken in context. The idea is not inconsistent with the philosophy behind distributional clustering methods. As in the work of Pereira *et al.*, a sense could be conceived of as a point in semantic space, the center of mass for a set of items determined according to similarity of context.

In sum, distributional methods show a great deal of promise for determining word classes automatically. However, they tend to conflate words with word senses, since corpora contain the former and not the latter. In addition, little attention has been paid to the semantics of the representations that result. In the next section I consider resolving these problems by adopting a "knowledge based" approach to distributional statistics.

2.4 Word Classes Based on a Taxonomy

A natural alternative to strictly distributional techniques for acquisition of lexical information, such as word classes, is the use of existing repositories of lexical knowledge, such as knowledge bases, lexical databases, dictionaries, and thesauruses. However, it is not always entirely clear how to make use of information from such sources in a statistical setting. There are several issues in particular that are worth mentioning.

- **Coverage.** In order to perform corpus-based analyses, adequate coverage of the corpus vocabulary is necessary. Traditional knowledge-based forms of "deep" lexical-conceptual knowledge — for example, the domain models in natural language systems like BBN's IRUS (Ayuso et al., 1989) — will require a great deal of effort if they are to scale up to large quantities of text, unless the domain is highly constrained. Text that is not constrained according to topic requires vocabulary coverage more on the order of that found in machine-readable dictionaries (MRDs).

- **Representation.** On the other hand, simply putting a dictionary into machine-readable form does not guarantee that it can be put to practical use as a computational tool. Numerous researchers have made progress in extracting computationally useful lexical information from MRDs and turning it into formal representations — e.g. (Alshawi, 1987; Byrd et al., 1987; Jensen and Binot, 1987; Braden-Harder and Zadrozny, 1991) — but some forms of information are easier to reliably extract than others. IS-A relationships between superordinate terms (hypernyms) and subordinate terms (hyponyms) appear relatively easy to discover (Byrd *et al.* report 98% accuracy in finding the head word in noun entries, taken to represent the noun's superordinate category), whereas other information may not be. Alshawi comments that "the fact that definition texts are often not analyzed completely means that information that is central to a definition is sometimes not taken into account" (p. 198).

- **Ambiguity.** Because most dictionaries and thesauruses relate words to other words, the automatic extraction of lexical information is to some extent confounded by lexical ambiguity. Byrd *et al.* comment that, absent an explicit indication of intended sense, some of their extraction methods are "relegated to the status of a semi-automatic (rather than a fully automatic) processing tool" (p. 234).

Given these considerations, it seems sensible from a practical standpoint to begin by investigating the statistical uses of word class information organized in the form of an IS-A taxonomy, specifically, a noun

taxonomy. For purposes of implementation, I will be using WordNet (Beckwith et al., 1991), a dictionary-sized, hand-constructed taxonomy of nouns in English. The use of WordNet's noun taxonomy addresses the issues of coverage and representation, and sense ambiguity is not a problem because it is organized according to noun senses that have been manually disambiguated.

Because reliable extraction of noun sense taxonomies from MRDs is quickly becoming a realistic goal, results achieved using WordNet will soon be more generally applicable. As techniques for extracting information from dictionaries continue to improve — and are applied to languages other than English — statistical methods developed using WordNet can be applied to the taxonomies that result. In addition to its practical advantages, WordNet is unlike most dictionaries by virtue of the goals with which it was designed: as a testbed for psycholinguistic principles of lexical organization. The theoretical foundations for WordNet, which I now discuss in some detail, will become relevant in Chapters 3 and 4.

2.4.1 The WordNet noun taxonomy and its semantics

WordNet (Miller, 1990b; Beckwith et al., 1991) is a large-scale resource for lexical information in English, constructed by George Miller and colleagues at Princeton University. It is broadly organized according to parts of speech — verbs, nouns, adjectives, and adverbs — and the information for words belonging to each of these syntactic classes is encoded in the form of a semantic network. Each network encodes different kinds of information; for example, the noun network contains links encoding relationships such as "part of" (CARBURETOR is a part of GASOLINE_ENGINE) and antonymy (EVIL is an antonym of GOOD), and the verb network encodes causal relationships (DISCLOSE causes BECOME_KNOWN) and manner distinctions (STRANGLE is a specialization of KILL).[10]

In this research, I have used only the noun taxonomy, and within that taxonomy, only two relationships: hyponymy and synonymy. Hyponymy as used by Miller *et al.* is intended to capture class inclusion, as in the classic example ELEPHANT IS-A MAMMAL. They write (Beckwith et al., 1991, p. 215):

> A noun X is said to be [a] hyponym of a noun Y if we can say that *X is a kind of Y*. This relation generates a hierarchical tree structure, i.e., a taxonomy. . . . A hyponym anywhere in the hierarchy can be said to be "a kind of" all of its superordinates. Researchers in artificial intelligence have long noted that a taxonomic organization is a highly efficient and economic storage system: all of the properties attributed to a superordinate node are inherited by its hyponyms; consequently, the properties need to be stored only once rather than separately with each hyponym.

A discussion in (Sparck Jones, 1964) on the definition of semantic relations points to a potential problem with hyponymy as Miller *et al.* are using it. Sparck Jones attempts to draw a distinction between "linguistic" relations, on the one hand, and "factual" relations, on the other — that is, lexical relationships based on interpretations of meaning as opposed to knowledge about the world.[11] The relationships between *woman* and *female* would seem to fall into the former category; the relationship between *woman* and *blouse* would seem to fall into the latter, since there is at most a contingent relationship having to do with the fact, in the world, that women sometimes wear blouses. Sparck Jones argues that the distinction is necessary in order to constrain the domain of lexical description: she comments (p. 48) that unless we do so,

[10] For purposes of notation, I use italics when referring to a word regardless of its sense, and uppercase when talking about a word in some specific sense. For example, one sense of *word* is in the sense of WORD or TIDINGS, as in "Have you received word?"

[11] Cf. Dowty's term *lexical entailment*, for which "the implication follows from the meaning of the predicate in question alone" (Dowty, 1991, p. 552).

> . . . we would finish up trying to give a description of the whole physical world, in the widest sense of "physical world." We would be constructing an encyclopedia, which is not what we want; and we could, moreover, never finish it (either medically or logically).

One definition of hyponymy that Sparck Jones considers is from (Lyons, 1961), where the lexical relationship is defined in terms of implication between sentences. In order for A to be a hyponym of B, it must be the case that a sentence containing A is understood in general to imply the same sentence with B substituted for A, but not conversely. So, for example, "x is scarlet" is generally understood as implying "x is red," though not conversely, and therefore *scarlet* is a hyponym of *red*. Implication is not used by Lyons as it is in formal logic; rather, it is intended to represent the judgements of the language user. As Sparck Jones describes it, "two sentences are equivalent if the ordinary language user would agree that in asserting the one we are asserting the other; and one sentence implies another if in saying the one we are prepared to say the other" (p. 54). I will distinguish implication of this kind from the normal use of implication in logic by calling it "plausible entailment."

A possible difficulty with using this definition in a theory of lexical semantics is that it fails to unequivocally rule out relationships that may be factual, as opposed to linguistic in the sense described above. So, for example, even if it were true that anything which is a dog is hairy, and that the ordinary language user would be prepared to say "x is hairy" given that "x is a dog," one might not want to say that that there is a hyponymic relation between the words *canine* and *hairy*. One the one hand, one could argue that being hairy is a definitional aspect of being a dog, in which case the relationship does belong within the domain of lexical semantics; but on the other hand, one could argue against this position, saying that the hairiness of dogs is accidental and therefore a matter of factual knowledge and not definition. The answer is not clear, and this, Sparck Jones argues, is the problem with trying to treat implication (plausible entailment) as a "linguistic" relation. She argues instead for making synonymy the central semantic relation in her lexical theory, a proposal that I will discuss momentarily.

Given this discussion, one could also conclude that the definition used by Miller *et al.* is not a linguistic characterization. Sparck Jones says as much:

> [We] might say that A and B are related [by hyponymy –PSR] if an a is a kind of b. But in this case we are obliged to say that we are no longer concerned with linguistic relations. (p. 63)

It is not entirely clear to me that the "kind of" relationship is in principle non-linguistic — after all, the central fact captured by such relationships is inheritance of properties, and even Sparck Jones would most likely be willing to concede that the inheritance of gender from WOMAN (female adult) to QUEEN (female monarch) is inextricably wrapped up in the meanings of the lexical items and not just accidental facts about the world. On the other hand, Miller *et al.* do not provide any rigorous definition of what they mean by "a kind of" — and, in fact, the contents of the noun taxonomy suggest that their criteria for superordinate-subordinate relationships are more generous than a strictly linguistic definition would allow.[12] For example, the classification of FRUIT (in its sense as foodstuff or produce, not part of a plant) as subordinate to GROCERIES (commodities sold by a grocer) surely reflects world or perhaps even culturally specific rather than linguistic knowledge.[13]

[12] All examples are from WordNet Version 1.2.

[13] Whether this is good or bad depends on your point of view. (Nirenburg and Raskin, 1987) argue that creating a conceptual model encoding "all the world's knowledge down to some level of detail" — what they call a World Concept Lexicon — is in fact a necessary preliminary to any construction of a lexicon for analysis or generation. However, they explicitly commit themselves to constrained domains, rather than the world in general, distinguishing themselves from broad-coverage knowledge acquisition endeavors such as CYC (Lenat, Prakash, and Shepherd, 1986).

Word Meanings	Word Forms				
	f_1	f_2	f_3	...	f_n
m_1		×			
m_2		×			×
⋮					
m_k			×		

Table 2.3: Schematic representation of a lexical matrix

The distinction between "linguistic" and "factual" knowledge is the source of quite a bit of difficulty; it is an issue that will arise again in Chapter 3. But assuming the distinction is applicable here, I will interpret hyponymy in the WordNet noun taxonomy as a lexical-conceptual relationship rather than as strictly "linguistic" — though the information is as strictly lexical as any one is likely to find in a dictionary, as distinguished from an encyclopedia.[14] From a formal point of view, the taxonomy does appear to respect implication in the sense used by Lyons: having said "x is a piece of fruit," the ordinary speaker of English can reasonably be expected to agree that "x is an item of groceries or foodstuffs" is also true, as well as any additional facts plausibly entailed by that statement on either linguistic or non-linguistic grounds. Thus, for the sake of making things concrete, I will formalize hyponymy in WordNet in the following way: if a and b are distinct hyponyms of c, and α, β, and γ are the sets of plausible entailments generated by virtue of membership in each respective class, then $\gamma \subset \alpha$, $\gamma \subset \beta$, and $\alpha \neq \beta$.[15]

The second organizing relationship within the WordNet noun taxonomy is synonymy. Miller *et al.* (1990) characterize synonymy in terms of a lexical matrix relating "word forms" to "word meanings" — such a matrix is represented schematically in Table 2.3. Word forms correspond to what I have earlier called "word tokens," and what Sparck Jones calls "word-signs" — essentially an atom or symbol, in written text typically a sequence of characters delimited by spaces.[16] Word meanings refer to "the lexicalized concept that a form can be used to express" (Miller et al., 1990, p. 4). However, an important facet of WordNet's formalization of word meaning is the absence — in both theoretical and practical terms — of a complete conceptual representation. Miller *et al.* distinguish between *constructive* theories and *differential* theories: the former requires representations that provide enough information to support accurate construction of the concept by person or machine, but the latter requires only that the representations of meaning are sufficient for someone to identify concepts that are already known. WordNet instantiates a differential theory by representing meanings in terms of *synonym sets* (sometimes shortened to *synsets*), so that m_2, for example, is identified by $\{f_2, f_n\}$, the members of its row in Table 2.3. In this example, m_2 might correspond to the

[14] Miller *et al.* comment that "somewhere a line must be drawn between lexical concepts and general knowledge, and WordNet is designed on the assumption that the standard lexicographic line is probably as distinct as any could be" (Miller et al., 1990, p. 15). See (Jackendoff, 1983, chapter 6) for an argument against taking the purported distinction between "semantic" and "conceptual" structure too seriously; Jackendoff's IS-INCLUDED-IN relation seems like it may also be a reasonable basis for hyponymy.

[15] It may ultimately be necessary to sort out the semantic issues here a bit more clearly, distinguishing implication, entailment, presupposition, and conventional implicature (see, e.g., discussion in (Dowty, Wall, and Peters, 1981)). Also see discussion of taxonomies and further references in Chapter 6 of (Cruse, 1986); in particular, Cruse makes a distinction between hyponymy and *taxonymy*, the latter being a sub-species of the former. A number of these issues arise in Chapter 3.

[16] In general, of course, some provision must be made for multi-word lexical items. That complication is ignored here.

concept BOARD in its sense as "a long, flat slab of sawed lumber" (AHD, 1991), with synonym set $\{f_2, f_n\}$ being $\{board, plank\}$. Miller *et al.* (1990, p. 4) comment:

> These synonym sets (synsets) do not explain what the concepts are; they merely signify that the concepts exist. People who know English are assumed to have already acquired the concepts, and are expected to recognize them from the words listed in the synset.

As it turns out, an approach to synonymy of very much the same kind is argued for at length by Sparck Jones, in the context of specifying which relationships are to be captured by an automatically constructed thesaurus. She makes a distinction similar to the one just described, and concludes (Sparck Jones, 1964, p. 28):

> The distinction between the two different ways of looking at a thesaurus head [i.e., a lexical grouping or class –PSR] is therefore a useful one, because it suggests a new approach to the problem of constructing a thesaurus; it suggests that we should treat a thesaurus head primarily as a set of words which are related to one another, and only secondarily as a set of words which express an idea. This means that we can set about the business of finding heads by looking for sets of words which are related in a suitable way, and then labelling them, rather than by inventing ideas, and then searching for words which express them. This approach has two advantages: the first is that the heads can be found more easily, and the second is that they can be found without any reference to an *a priori* set of ideas.

Sparck Jones characterizes synonymy as follows. First, as a background assumption sentences are taken to have a property she calls a *ploy* (i.e. the way in which it is em*ploy*ed) — for example, *Shut up* and *Keep quiet* have the same ploy. Given a sentence S and one of the word positions in S, a *row* is defined as a set of words that can appear in that position without changing the ploy of S. So, for example, the sentences in (3) give rise to the row $\{signal, sign\}$ and (4) gives rise to the row $\{shouted, cried, called\}$.

(3) a. He gave the signal for the advance.
    b. He gave the sign for the advance.

(4) a. He shouted for help.
    b. He cried for help.
    c. He called for help.

One might argue that, even if ploy can be given some precise interpretation, no two words are ever truly substitutable even in a particular context, since each carries different associations or overtones. Sparck Jones responds to this objection as follows (p. 84):

> In spite of this argument, it is an empirical fact that we do explain the meaning of a word in a context by giving other words which we say can be used in the same way; we do in practice say that words may be used synonymously. This suggests that we can make a distinction between a particular use and the whole range of uses of a word. We can and do say that though the overtones of two words, representing their whole ranges of uses may be different, their uses in a particular context may, for all practical purposes, be treated as indistinguishable.

Given this discussion, it seems clear that Table 2.3 is representative of the theories of both Miller *et al.* and Sparck Jones. In both cases, the goal is not to provide a thesaurus defined in terms of word synonymy;[17] rather,

[17] Indeed, Sparck Jones expends considerable effort detailing the arguments against this view, and later proposes that "total" synonymy be derived from synonymy of word uses.

taxonomic classes comprise differential representations corresponding to word uses that are synonymous in particular contexts. Although WordNet classes represent a higher level of abstraction than Sparck Jones's rows — something like what she calls a *group* — the core idea is the same. Lexical classes comprise sets of words that in some contexts are mutually substitutable while preserving the essential components of meaning and use.

To summarize, one way to characterize synonymy in WordNet is as follows. Let s be a WordNet synset containing synonyms (word forms) $\{w_i\}$. Then there is a "representative" set of sentences $\{S_j\}$ such that if S_j entails σ, then $S_j[n = w]$ also entails σ for all $w \in s$, where $S_j[n = w]$ denotes sentence S_j with word w substituted at position n; furthermore, s contains all such w. So, for example, suppose s is the synset $\{board, plank\}$: for all practical purposes the sentences in (5) are interchangeable, in that the conclusions one can draw — such as σ_1 and σ_2 — are the same.

(5) a. John sawed the board in two.
b. John sawed the plank in two.

(6) a. $[\sigma_1]$ The thing John sawed is used for construction.
b. $[\sigma_2]$ The thing John sawed is flat.

Furthermore, some entailment serves to exclude other similar words from the set; for example, *brick* and *shingle* are closely related to this sense of *board*, but substituting *brick* for *board* in (5a) would violate σ_2, and substituting *shingle* would introduce a new entailment σ_3 not shared by *board* and *plank*.

(7) $[\sigma_3]$ The thing John sawed is used for covering roofs.

Again, it should be noted that the notion of entailment that is operative here is concerned not with logical necessity, but with implication in the sense used by (Lyons, 1961) — that is, entailed properties consist in conclusions that one would be willing to draw. It is not difficult to elicit such properties from human subjects; for example, (McRae, de Sa, and Seidenberg, 1992) describe a norming study in which subjects produced properties like *requires a driver* and *used for transportation* (e.g., given *bus*). Although I have characterized relationships in WordNet in terms of such properties, however, it is important to remember that this is done solely in the interest of providing a well-founded interpretation for an existing taxonomy — such relationships are not explicitly represented in WordNet, nor were they explicitly used in its construction.

Computationally, these two taxonomic relationships in WordNet, hyponymy and synonymy, form the basis for what is essentially an inheritance system. Each word token w is mapped to senses(w), a set of synonym sets that in effect represent all of its word senses. For the sake of notational convenience, such synonym sets will be represented not by listing each included word, but by pairing a single descriptive word with a unique identifier — for example, ⟨`board,4012740`⟩ rather than $\{board, plank\}$.[18] Each synset s in the taxonomy may have hyponyms (subordinates) and hypernyms (superordinates); these correspond to both directions of an IS-A link. Because WordNet permits multiple inheritance, hyponymy and hypernymy are one-to-many relationships, and the structure of the taxonomy is a directed acyclic graph rather than a tree.

[18] The numerical identifiers are derived from WordNet's internal data representation and can be thought of as arbitrary; hence this is equivalent to providing traditional sense labels like BOARD1, BOARD2, etc. I will sometimes omit the unique identifier when it is not particularly relevant, e.g. writing ⟨`board`⟩ rather than ⟨`board,4012740`⟩. Identifiers used here are from WordNet version 1.2.

Description	Example	Count
Hyphenated term	grease-removal	17
Occurred once in corpus	carousing	16
Number	.05%	4
Not in dictionary	fella	3
Unusual or wrong spelling	threshhold	2
Acronym	USP	1
Other	alkali	6

Table 2.4: Brown corpus nouns missing from WordNet

Each synset in WordNet can be viewed as a class containing all the words in all directly or indirectly subordinate synsets — that is, all synonym sets that inherit its plausible entailments. The extensional interpretation of a class c will be written using the notation words(c); for example, synset $\langle$`lumber,4012560`$\rangle$, interpreted as an extensional class, contains *batten*, *board*, *deal*, *fin*, *furring*, *lath*, *louver*, *louvre*, *lumber*, *pale*, *picket*, *plank*, *slat*, *spline*, *stave*, *strip*, and *timber*.[19] Conversely, classes(w) will represent the set $\{c|w \in \text{words}(c)\}$ — notice that this includes *all* the classes in which word token w is contained, regardless of whether a particular sense of w was intended. From this point on, notation that is ambiguous between synonym sets and classes — e.g. $\langle$`lumber,4012560`$\rangle$ — should be interpreted as referring to the class, unless otherwise noted.

In terms of size, the WordNet noun taxonomy (version 1.2) contains on the order of 35,000 synonym sets, and a vocabulary of on the order of 47,000 nouns (approximately 30,000 if compounds are excluded). Using the Brown corpus as a representative sample of English, WordNet's coverage accounts for approximately 95% of the tokens tagged as common nouns (singular or plural). Table 2.4 shows a breakdown into categories for 49 nouns randomly chosen from the set of nouns in the Brown Corpus that do not appear in the WordNet taxonomy. The reference dictionary for the "not in dictionary" category was (AHD, 1991), and the one acronym on the list, USP, derives from the United States Pharmacopoeia reference standard. Of the hyphenated terms, just one, *vice-president*, appears in WordNet as a compound.

This coverage estimate will most likely remain fairly stable as WordNet changes: of the six nouns in the "other" category — *alkali*, *glycol*, *growl*, *handspike*, *quicksilver*, and *slam* — only one, *alkali*, appears in the more recent WordNet Version 1.3.

2.4.2 Estimation of class probabilities

Although the discussion in Section 2.4.1 describes a structured taxonomy, the probabilistic formalization presented here will be based on a sample space consisting of class labels. In essence, the theoretical formulation of the sample space will suppose that people say things like "I drank some $\langle$`beverage`$\rangle$" rather than "I drank some wine." The structure of the taxonomy will play a role not in defining the sample space, but in estimating probabilities.

[19] Actually, it also contains the compound noun *furring_strip*; however, I treat all compound nouns in WordNet as if they did not exist. Although this potentially loses useful information, it does away with the problem of determining whether or not a nominal compound should be treated as a unit.

To be specific, let the probability space $\langle \Omega, \mathcal{F}, \mathrm{p} \rangle$ consist of:

$$\begin{aligned} \Omega &= \{c_1, c_2, \ldots, c_k\} \\ \mathcal{F} &= \mathcal{P}(\Omega) \\ \mathrm{p} &: \mathcal{F} \rightarrow [0, 1], \end{aligned}$$

where Ω is the complete set of unique class identifiers in the taxonomy. $\mathcal{F}$ is simply the maximally fine-grained event space based on such a sample space, and p is a probability function. Thus the probability space here is just like the space for rolling a die; in this case it just happens that the die has k sides.

Since real observations contain not class labels but words, the frequency of a class c will be estimated as

$$\mathrm{freq}(c) = \sum_{w \in words(c)} \frac{1}{|classes(w)|} \mathrm{freq}(w). \tag{2.12}$$

Whenever word w is observed, credit must be assigned to some of the classes in the sample space. So, for example, if the actual observation is *drink wine*, the frequency of co-occurrence with *drink* will be incremented for $\langle$`wine,2657055`$\rangle$, $\langle$`alcoholic_beverage,2654808`$\rangle$, $\langle$`beverage,2653465`$\rangle$, and $\langle$`5941,substance`$\rangle$, among others. At the moment they are incremented by equal amounts, something I will discuss further in detail in a moment.

There are many ways to estimate class probabilities $p(c)$ on the basis of such a frequency distribution, of which the maximum likelihood estimate (MLE) is the simplest. Although there are known problems with maximum likelihood estimates of probability (see discussion in (Church and Gale, 19xx)), MLE seems a reasonable starting point, especially since one of its main problems — the assignment of zero probability to all unseen data — is in fact one of the problems this work is attempting to resolve. A similar point is made by Pereira *et al.* (1993), who write, "We could [smooth zero frequencies] . . . However, this is not very satisfactory because one of the goals of our work is precisely to avoid the problems of data sparseness." Results obtained using MLE should be further improved by using improved probability estimates.[20]

The estimation of probabilities using MLE is straightforward:

$$\hat{p}_{MLE}(c) = \frac{\mathrm{freq}(c)}{N}, \tag{2.13}$$

where $N = \sum_{c'} \mathrm{freq}(c')$. The calculation of joint probabilities is similarly straightforward: if the sample contains co-occurrences (x, w), where x is an element of some set $\mathcal{X}$ of tokens, one need only replace equation (2.12) with

$$\mathrm{freq}(x, c) = \sum_{w \in words(c)} \frac{1}{|classes(w)|} \mathrm{freq}(x, w). \tag{2.14}$$

For example, such a sample might contain pairs consisting of a verb and its direct object, in which case $\hat{p}_{MLE}(v, c)$ would represent the estimated probability of a member of c appearing as the direct object of v.

Let us return to the fact that in equation (2.12) the observation of a word w has an equal effect on every class to which w belongs. This is clearly an oversimplification. However, absent a solution to the problem of word sense disambiguation, distributing the "credit" for a word uniformly over its possible classes seems the most sensible solution.

This brute-force approach works because related words tend to be ambiguous in different ways. For example, consider the observation of two verb-object combinations, *drink wine* and *drink water*. On the basis

[20] In earlier work, I used Good-Turing estimates rather than MLE; for discussion of this and further notes on probability estimation, see Appendix A.

of these observations, the joint frequency will be incremented for each class containing *wine* in any sense — including, for example, ⟨chromatic_color,1925370⟩. Similarly, the second pair will be recorded as a co-occurrence between *drink* and inappropriate categorizations such as ⟨body_of_water,2995307⟩. However, evidence for co-occurrence will *accumulate* only for classes containing both *water* and *wine*, such as ⟨beverage,2653465⟩. The cumulative evidence thus will tend to support appropriate interpretations, and counts with inappropriate senses will appear only as very low frequencies dispersed throughout the taxonomy. A similar point is made by Yarowsky (1992), commenting on the calculation of statistics using the numbered categories in Roget's Thesaurus:

> While the level of noise introduced through polysemy is substantial, it can usually be tolerated because the spurious senses are distributed through the 1041 other categories, whereas the signal is concentrated in just one. (p. 455)

Given the much larger set of categories in WordNet, the dispersal of inappropriate senses should be even more effective. However, using classes at all levels of the WordNet taxonomy has its disadvantages, relative to the flat set of Roget's categories used by Yarowsky: classes low in the taxonomy accumulate less evidence than classes higher up. As a result, among small classes it can be more difficult distinguish which correlations are signal and which are noise. For example, given numerous observations of verb *blow* with object *nose*, there is not enough accumulation of evidence to determine that the high frequency with ⟨nose,2088032⟩ (the body part) is appropriate but the high frequency for ⟨nose,1172320⟩ (e.g. the front part of an aircraft) is not.

2.4.3 Comparison with distributional methods

The following discussion is organized around several of the issues that have been raised in this section.

Sparseness. Knowledge-based taxonomies and distributional methods both address the problem of sparse data. Knowledge-based methods have the advantage of classifying words that have not been encountered at all, or words that are difficult to classify distributionally owing to lack of evidence. However, every dictionary has gaps, and unlike distributional methods, knowledge-based taxonomies are not well-suited to keeping up with changing terminology, proper names, and productive variations in usage. Hearst and Schütze (1993) present a promising approach to this problem; their work suggests that distributional methods can be used to classify new words and proper names within the WordNet taxonomy.

Abstraction. Taxonomies of the kind defined in Section 2.4.1 provide a rich source of information about lexical items, expressed at a level of abstraction that seems suitable for capturing conceptual in addition to simply lexical information. However, some relationships may be more genuinely lexical than conceptual — for example, (Smadja, 1991) argues that the distinction between *strong tea* and *powerful tea* cannot be accounted for on purely syntactic or semantic grounds, and thus should be considered an instance of lexical idiosyncracy. Capturing such relationships may be more difficult using a taxonomy like WordNet as an intermediary. Therefore, consistent with the approach taken by Hearst and Schütze, the use of a word sense taxonomy should be viewed as a way to extend, not replace, purely lexical methods.[21]

[21] One could argue that some relationships seem "purely" lexical simply because the appropriate syntactic or semantic characterization has not yet been found — for example, Patrick Hanks has conjectured that *strong* describes an intrinsic property whereas *powerful*

Computational expense. The distributional derivation of classes clearly entails a great deal of computational expense. Such expense can be justified on the grounds that the cost will vanish in the limit as computers become more and more powerful; however, the sheer combinatorics of lexical relationships suggests that some forms of distributional analysis may never be tractable. For the time being, even highly optimized algorithms, such as those presented by (Brown et al., 1992), require a great deal of computation.

On the other hand, dictionary building requires an enormous amount of effort — work undertaken by people, not machines. Although it is increasingly possible to take advantage of existing dictionary resources that appear in machine-readable form, the extraction of useful information from such sources — for example, by parsing dictionary entries — can itself require significant computation. Moreover, there is no guarantee that existing sources of information, constructed as they are for other purposes, can be used for computational purposes without substantial modification, if at all.

Perhaps a more interesting issue is the cost of keeping up with language use as it changes over time. Most distributional methods involve global computations, requiring a complete recomputation of the model even if the new version will differ significantly from the old only in a small fraction of cases. In contrast, on-line knowledge sources can be modified incrementally without affecting core information that is retained over time.

Class identifiers. There are two reasons for associating symbolic descriptions with word classes: human readability, and integration with other symbol-based system components. To some extent, WordNet comes with human-readable class labels built in — this property will be shared by any knowledge base constructed according to what Miller terms a differential theory of representation (see discussion in Section 2.4.1).[22] Distributionally-derived classes may or may not be similarly interpretable; for example, showing the three or four words closest to the cluster centroid, as done by Pereira *et al.* (1993), leads to some easy interpretations (e.g. the centroid for *recognition*, *acclaim*, *renown*, and *nomination*) and some that are rather more opaque (e.g. *pollution*, *failure*, *increase*, and *infection*).

Symbolic class labels can also be necessary in order to integrate word classes with other components of a language-processing system — for example, once a sense disambiguation algorithm has selected the most likely category for a word in context, the rest of a language-processing system will need to make use of other information indexed by that category. Such an integration may be more difficult when the only representation of classes is extensional. On the other hand, class labels alone do not suffice: it is also necessary to define a mapping between the set of class labels and the set of symbols used by other components of the system. At one extreme, the other components use the class taxonomy as a reference point and this task is trivial; at the other extreme, the other components have their own distinct characterization of word categorization and the problem effectively becomes one of merging ontologies (Knight, 1993).

In some cases, neither human readability nor system integration is an issue and word equivalence classes can be used as such without further interpretation. This is most notably the case for class-based language models — although even in such cases the ability to describe classes concisely may help, for example in interpreting what exactly the language model has done when it makes an error.

---

describes an extrinsic property (Church and Mercer, 1993, p. 20). This issue is not likely to be resolved any time soon, given our current limited understanding of lexical semantics, so maintaining Smadja's distinction seems reasonable.

[22] Even WordNet has exceptions, of course; for example, synonym sets containing only a single synonym. Peter Norvig (personal communication) has explored methods for rendering synsets in human-readable form — for example, augmenting a singleton synonym set with terms from its immediate superclass, e.g. distinguishing FILE/RECORD vs. FILE/HAND_TOOL.

The semantics of word classes. Although, as discussed in Section 2.4.1, it is difficult to ground taxonomic representations such as WordNet in precise formal terms, the use of the WordNet taxonomy makes reasonably clear the nature of the relationships being represented. (The same may not be true of other taxonomies — see for instance Sparck Jones's discussion of the various relationships that relate words in Roget's Thesaurus.) However, like the need for class identifiers, the extent to which the contents of a class need to be clearly describable will vary depending on the problem being solved. For tasks such as topic labelling, a group of words that are associated according to topic may suffice (Hearst and Schütze, 1993); for integration of classes within a language understanding system framework, the ability to more clearly identify what a class contains may be important.

Conflation of word senses. The distributional hypothesis — that words sharing distributional contexts will be similar to some degree — makes the most sense when taken at the level of word meanings or uses rather than word tokens. Consider a single polysemous word token: its distributional signature may capture the essential connection between a word's different uses (e.g. a *door* as an entranceway as well as the barrier that occupies the entranceway), or the distribution may sum together quite independent uses (e.g. *crane* as bird versus machine), or it may fall somewhere in between. In the first case, the distributional hypothesis clearly holds. In the second case, the hypothesis runs into trouble, although perhaps the distribution can somehow be analyzed into its independent components, and those used as the basis for judging similarity. (This would amount to extracting easily distinguished word senses from the distribution of the word token.) The third case is the most problematic: two word senses that intuitively are distinct will be treated as one on the basis of the word token's distribution.

Although it is possible to induce word senses as entities in their own right on the basis of word token observations, as in (Pereira, Tishby, and Lee, 1993), the discussion in Section 2.3.4 suggests that capturing word similarity will ultimately require a distributional analysis of word senses rather than word tokens.[23] Such a solution is circular, however, under the assumption that the set of word senses is itself defined by analyzing how word tokens are distributed. The proposal here to use externally-defined word senses can be seen as one way of breaking that circularity: given a distributional analysis of classes using the model of Section 2.4.2, it becomes possible to consider both the distributional behavior of words and the distributional behavior of the classes to which they belong.[24]

Distinction between domain-dependent and domain-independent knowledge. Information extracted from a corpus is always domain dependent to some extent, even for corpora that aspire to balanced coverage. Church and Mercer (1993, p. 19) argue that if corpora are combined to obtain larger quantities of data, the quirks of individual subcorpora can essentially be treated as noise:

> Fortunately, though, it is extremely unlikely that [phrases specific to an individual corpus] will appear much more often than chance across a range of other corpora . . . If such a phrase were to appear relatively often across a range of such diverse corpora, then it is probably worthy of further investigation. Thus, it is not required that the corpora be balanced, but rather that their quirks be uncorrelated across a range of different corpora. This is a much weaker and

[23] See especially equations (2.10) and (2.11) in Section 2.3.4.

[24] Re-estimation is another possible way out of this circularity, though I will not pursue that idea further here.

> more realistic requirement than the more standard (and more idealistic) practice of balancing and purging quirks.

In practice, however, statistical methods are often useful for capturing language as it is found within a particular context — automated tailoring to a particular domain is, in fact, one of the biggest advantages of corpus-based techniques. If the goal is to parse IBM manuals, for example, then more general language usage in sources like the Brown corpus may be misleading: things that get installed in general (8b) may not be the kinds of things that get installed when dealing with computers (8a).

(8) a. After installation of the option, the backup copy of the Reference Diskette was started for the computer to automatically configure itself.
b. Makes necessary purchases, places contracts, supervises construction, installation, finishing and placement of furniture, fixtures and other correlated furnishings . . .

Knowledge-based sources of information also have their quirks. However, by using taxonomic classes rather than distributionally derived classes, it is possible to isolate word classification from other distributional issues. As a result, the behavior of class-based statistical methods can be broken down according to corpus-dependent and corpus-independent factors. This should be useful for identifying differences between corpora as well as for evaluating the quality of the taxonomy itself.

In conclusion, distributional methods and knowledge-based methods for using word classes each have advantages and disadvantages. Quite a few researchers have started investigating the derivation of word classes on the basis of distributional similarity; in this chapter I have attempted to provide a balanced description of their approaches. The present work differs from those investigations, beginning instead with existing large-scale knowledge sources as a source of lexical information.

It seems clear that although lexical taxonomies cannot provide all the knowledge one could hope to extract from a corpus, neither are distributional methods likely to capture all the information found in a resource like WordNet, even in principle. Methodologically, the use of classes from a taxonomy makes it possible to be clear about what kind of knowledge is represented, to keep separate what information is provided by the classes and what by the lexical statistics, and to take advantage of existing resources.

Finally, an underlying goal of this work has been to incorporate knowledge-based information in to statistical methods in the most conservative fashion possible. Although there is always some residual uncertainty as to how to interpret notions like synonymy and hyponymy — at least in terms of formal semantics — the synonym and IS-A relationships captured within the WordNet taxonomy are intuitively reasonable and as widely accepted as any other form of knowledge representation. Furthermore, even those representational details can be separated from the way the taxonomy is used in determining probabilities based on classes: the class frequency estimate proposed here assumes only an extensional representation of classes as sets of words, without requiring any further interpretation. The methods developed here can therefore be applied in other settings, where some other model determines the criterion for coherent sets of words.

In the next chapter I turn to selectional constraints. After providing a review of the philosophical and linguistic issues, I will argue that previous approaches leave unresolved problems concerning how such constraints are to be formalized. These problems, I will suggest, can be resolved in part by making use of

the taxonomic lexical representation discussed here, and in part by formalizing selectional constraints using a model within which class-based probabilities play a vital role.

Chapter 3

An Information-Theoretic Account of Selectional Constraints

Limits on the applicability of predicates to arguments have variously been called sortal constraints, selection (or selectional) restrictions or constraints, and type rules. A review of the literature on such constraints suggests two different ways in which they can be characterized: a "semantic" approach, which has been questioned on empirical grounds, and an "inferential" approach, for which the theoretical issues are at present poorly understood. In this chapter, I propose a new, information-theoretic formalization of selectional constraints, and argue that it addresses both theoretical and empirical concerns.

3.1 Overview

The topic of this chapter can be traced back to Aristotle.[1] He considered the case of sentences like

(9) a. α is even.
b. α is not even.

when α is not a number — for example, when α is *Socrates*. Intuitively, "Socrates is even" is certainly not true, but neither is it clearly false; rather, it seems to be a case where α is simply not the sort of thing to which the predicate can be applied. More recent examples of this phenomenon include Chomsky's predications concerning ideas and sincerity:

(10) a. Colorless green ideas sleep furiously.
b. Sincerity may admire the boy.

Limitations on the applicability of a predicate have variously been called sortal constraints, selection (or selectional) restrictions or constraints, and type rules; expressions violating constraints of this kind have been discussed in terms of category mistakes, selectional violations, type crossings, and semantic anomaly. Whatever the guise, phenomena of this kind raise a number of difficult issues.

[1] I regret that I am not nearly as erudite as this opening statement might suggest: in this section and the one that follows I draw heavily on the deep and insightful discussion found in (Horn, 1989), especially chapters 2 and 6.

First, examples like (9) pose a problem for straightforward truth-theoretic semantics: the axiom that every proposition is either true or false (Aristotle's "law of the excluded middle") confronts the widely-held intuition that sentences like "Socrates is even" are neither true nor false, but instead nonsensical, absurd, or meaningless.

Second, in addition to the philosophical problems posed by anomalous sentences of this kind, it is necessary to consider the positive application of predicate-argument constraints in the process of interpretation. Someone reading (9) — under the mathematical interpretation of *even* — will infer that α is an integer. Similarly, to take an example from (Fodor, 1977, p. 195), the pronoun in

(11) This one admires John.

will have attributed to it properties consistent with its use as the subject of *admire*, such as animacy and the possession of higher psychological functions.

A third, closely related point concerns the interpretation of ambiguous lexical items. To take an example from (McCawley, 1968, p. 131),

(12) a. John has memorized the score of the Ninth Symphony.
b. The score of the Ninth Symphony is lying on the piano.

lexical items that denote works of art or scholarship can also denote the physical embodiment of those works; example (12) suggests that the correct interpretation of *score* may in part derive from constraints provided by the verb.

I consider the first of these issues in Section 3.2, where I review some of the philosophical approaches to sentences like "Socrates is (not) even" and recapitulate Horn's (1989) argument for a pragmatic rather than semantic treatment of (negative) category mistakes. In Section 3.3, I turn to the characterization of selection restrictions in generative linguistics, particularly the discussions of Katz and Fodor (1964) and McCawley (1968), who concern themselves not only with selectional violations but also with the role of selection restrictions in constraining the interpretation of non-anomalous sentences (issues two and three).

At this point, I will suggest that the discussion has reached something of an impasse (Section 3.4). On the one hand, it will have become evident that the truth-theoretic approaches to selection restrictions — that is, most of the voluminous literature reviewed by Horn — rest on the assumption that such restrictions are phrased in terms of necessary and sufficient semantic conditions on the applicability of a predicate. I will observe that this approach is equivalent to the "defining properties" approach to mental categories and thus inherits some substantial empirical problems associated with that approach. On the other hand, I will argue that the "pragmatic" or "inferential" approach proposed by Horn (aspects of which can be found in McCawley and in (Drange, 1966) and (Johnson-Laird, 1983)) is too open-ended, potentially requiring a theory that encompasses all the inferences a person might make on the basis of factual knowledge.

The rest of the chapter, naturally enough, will be devoted to a proposal for going beyond the impasse. In Section 3.5, I will focus on the inferential view, but argue that, rather than incorporating all kinds of knowledge and inference, a characterization of selectional phenomena can be formulated in terms of an extremely restrictive knowledge representation, together with an appropriate information-theoretic characterization of predicate-argument relationships. By replacing selectional *restrictions* with an information-theoretic proposal for selectional *preference*, I provide a precise model that is nonetheless consistent with the inferential

view of the verification and transfer of properties in anomalous and non-anomalous utterances. In Section 3.7, I describe a computational implementation of this model, and illustrate its behavior on examples from throughout the chapter. In addition, I consider the possibility that this account might provide the groundwork for a psycholinguistic model of "local semantic fit" and thus contribute to the study of on-line plausibility effects in sentence processing. I conclude in Section 3.8 with a brief consideration of the present approach in relation to other computational proposals for capturing selectional constraints.

3.2 Category Mistakes

In this section, I rely to a great extent on the discussion in (Horn, 1989), who describes utterances such as (13) as *category mistakes* (CMs):

(13) a. The number two is blue.
b. The number two is not blue.

The latter sentence is an instance of a *negative* category mistake (NCM).

These and closely related expressions have generated a formidable quantity of philosophical discussion — at one point Horn refers to "standard approaches of dual-negation logics" in which (13b) can be read as "simply false, a priori false, neither true nor false, false and insecure, or meaningless, depending on whether the assessor is, respectively, an Aristotelian, a Drangean, a Bochvarian, a Bergmannian, or a Russellian" (p. 139). On the following page he presents a chart in which no fewer than twenty-two other philosophers join the five just cited.

Rather than attempting the hopeless task of improving on Horn's review of the literature, I will restrict myself to a summary of the major issues that are relevant to the present study. These are the questions of whether category mistakes are best analyzed in terms of entailments or presuppositions, whether they should be considered meaningful or meaningless, and whether they should be treated within a theory of (truth-conditional) meaning or from a broader, pragmatic perspective.

3.2.1 Entailment

If predicate `P` does not "naturally" apply to argument `x`, as is the case with `blue` and the number two, what can be said about the truth value of `P(x)`? One might begin by supposing that any predicate has associated with it a domain of applicability, and that applicability is an *entailment* of the predication. On this view, introduced by Aristotle, the predication (13a) entails that the number two be something that can be described as having a color, and since this entailment is false, (13a) simply is false as well.[2]

Sentence (13b) could be described as generating the same entailment, and therefore also false. At first glance, this would seem to be problematic, since it suggests that both a proposition and its negation could be simultaneously false. However, it is not necessarily the case that (13b) is the negation of (13a). Aristotle's analysis of (13b) posits two different readings: in one the entire *predicate* is negated, and in the other what is negated is the term `blue` itself. These interpretations can respectively be phrased as follows:

(14) a. The number two is-not blue.

[2]For the moment, nouns are to be interpreted literally — so although "the number two" might be used to refer to, say, a sheet of paper cut out in the appropriate shape, here it should be interpreted as referring to the abstract mathematical concept. The ability to *construe* arguments on other than their intended readings will be discussed in Section 3.5.

b. The number two is not-blue.

On this analysis, the former interpretation does not ascribe a color to the number two, but merely negates the predication (13a) that does so — therefore applicability of the predicate is not entailed and the statement is simply true (since the number two is not among the set of blue things). The latter reading is an instance of *term negation*, in this case ascribing the property `not blue` to the number two. Unlike predicate negation, term negation requires that the applicability of a property be considered; in essence, `not blue` is to be interpreted as equivalent to `red or green or yellow or...`. As a result, (13b) *does* entail the applicability of color terms to the argument, and is therefore false.

Horn points out that this analysis of category mistakes is closely tied to the analysis of sentences containing empty (i.e. non-denoting) subjects, such as (15):

(15) a. The king of France is bald.
b. The king of France is not bald.

Taking Aristotle's view, sentence (15b) can be treated as an instance of term negation — affirmation of the predicate `not bald` of the king of France — which entails that the king of France exist. Alternatively, it can be treated as an instance of predicate negation, which does not generate such an entailment.

3.2.2 Presupposition and meaninglessness

A widely discussed alternative to the entailment approach, associated in particular with Strawson, is that statements do not *entail* the applicability of predicates to subjects (or the existence of subjects), but rather that applicability (and existence) are *presupposed*. On this view, someone uttering (15a) has not asserted that a unique king of France exists, but rather has acted on the presupposition that such a person exists. If the presupposition is not true, then the truth value of (15a) is simply not at issue; it is neither true nor false.

It is important to distinguish between a proposition being neither true nor false and its being *meaningless*. An uncontroversially meaningless expression is (16):

(16) Boy girl of picture saw the and.

"Word salad" sentences like this result in more word salad when embedded in a matrix sentence:

(17) I dreamed that boy girl of picture saw the and.

and there is no intuition that the import of (17), if there were one, could be distinguished from another instance of the same kind of anomaly:

(18) I dreamed that the saw picture of girl boy and.

Sentences that appear more normal on the surface can nonetheless exhibit behavior similar to this; for example, violations of Grimshaw's (1979) s-selection:

(19) a. John believed if the train left on time.
b. Mary reported that John believed if the train left on time.

In contrast, category mistakes can be felicitously embedded (20) and distinguished from each other, since in (21a,b) Mary is clearly making two different claims:

(20) I dreamed that the number two was blue.

(21) a. Mary claims that colorless green ideas sleep furiously.
b. Mary claims that sincerity may admire John.

Furthermore, category mistakes can be entailed by and entail other propositions. For example,

(22) Quadratic equations do not go to race meetings.

entails

(23) Quadratic equations do not watch the Newmarket horse races.

and is entailed by

(24) Quadratic equations do not move in space.

Similar observations hold of "empty subject" cases like those involving the king of France.

Propositions that generate entailments can certainly not be viewed as meaningless. One way of describing the truth-conditional status of category mistakes, then, is to say that they are meaningful, but in some sense *insignificant*. "Presuppositionalist" accounts generally accomplish this by relaxing the assumption that the truth function be complete, instead characterizing it as a partial function and permitting category mistakes to fall into a "truth value gap," or by using a multi-valued logic, with category mistakes taking an intermediate value expressing something like "not true." Crucially, in such a system the non-truth of a proposition will *not* entail the truth of its negation.[3]

3.2.3 Implicatures, pragmatics, and metalinguistic negation

To summarize thus far, one treatment of anomalous expressions involving inappropriate or empty subjects retains a truth-theoretic analysis in which every proposition is either true or false; on this view the existence and appropriateness of the subject amounts to an entailment of the proposition. Another widely held view analyzes these not as entailments but as presuppositions, requiring some semantic status other than truth or falsity.

Presuppositional treatments are complicated by the fact that negative category mistakes and negative existentially-presupposing expressions have a reading in which the presupposition appears to be "cancelled":

(25) a. The king of France isn't bald — there *is* no king of France!
b. Ideas aren't green — they're only in your head![4]

In these cases, it is evident that the speaker is asserting the truth of the first statement, and justifying the assertion by explicitly *rejecting* a presupposition.

[3] For a presuppositional treatment of category mistakes within the framework of Montague semantics, see (Waldo, 1979), who follows van Fraasen's (1968) method of supervaluations in treating truth as a partial function. Interestingly, Waldo's method succeeds in evaluating "The theory of relativity is shiny" as neither true nor false, while still managing to evaluate "Every shiny theory of relativity is shiny" as a tautology; however, no account is given for category mistakes that appear as the embedded clause in embedded contexts.

[4] This example is adapted from an exchange between Barbara Landau and a blind five-year-old subject, brought to my attention by Lila Gleitman:

(a) Barbara: Could an idea be green?

(b) Blind child: No, silly! They're only in your head.

An alternative that is still very much in the spirit of presuppositionalist approaches, proposed by Karttunen and Peters (1979), distinguishes between the truth-conditional semantics of a proposition and its *conventional implicatures*. Conventional implicatures are distinguished from *conversational* implicatures in that in the former case "the conventional meaning of the words," as opposed to general conversational principles, will determine what is implicated (Grice, 1975, p. 66). Karttunen and Peters focus on cases like the following:

(26) a. John managed to solve the problem.
b. John didn't manage to solve the problem.

(27) It was difficult for John to solve the problem.

(28) John solved the problem.

Here the implicatum (27) is part of the meaning of both the positive and the negative statements in (26), but not part of the truth-conditional meaning, since (26a) and (28) are truth-conditionally equivalent. The distinction is formalized by breaking down the meaning φ of an utterance into a pair $\langle \varphi^e, \varphi^i \rangle$, respectively representing the utterance's truth-conditional meaning and its conventional implicata.

The ambiguity between presupposition-preserving and presupposition-cancelling negation is then accounted for by the existence of two negation operators:

$\langle \neg\varphi^e, \varphi^i \rangle$ [Ordinary negation of φ]

$\langle \neg(\varphi^e \wedge \varphi^i), (\varphi^i \vee \neg\varphi^i) \rangle$ [Contradiction negation of φ]

Ordinary negation affects only the truth-conditional meaning, preserving implicatures (roughly equivalent to presuppositions), so that John's not managing to solve the problem still implicates that the problem was difficult. In contrast, the meaning of the "contradiction" negation leaves unspecified whether the negation is based on the truth-conditional meaning or the conventional implicata; and crucially, the implicata of the contradictory negation are vacuous, thus in effect cancelled.

Although Karttunen and Peters do not discuss category mistakes explicitly, it is not difficult to see how directly this approach would adapt to such cases: applicability conditions of predicates (e.g. having a physical surface for color predicates like `green`) can be taken as conventional implicatures. As a result, the utterance "Ideas are not green" has an interpretation as an ordinary negation, in which case it suffers from the same presupposition failure as "Ideas are green," or as a contradiction negation, as in example (25b). Notice the analogy between this approach and the term- vs. predicate-negation ambiguity discussed earlier.

Horn (p. 146) comments that "as part of the meaning of an expression and yet not part of its literal meaning (that aspect of meaning which affects truth and satisfaction), conventional implicata are located simultaneously within semantics . . . and pragmatics." However, the ambiguity of negation is still a semantic ambiguity since it concerns only meaning, albeit construed slightly more broadly. Having gone this far, he argues in favor of going still further, proposing an analysis according to which the ambiguity of negation unashamedly involves pragmatics.

At the core of Horn's argument is a distinction between *descriptive* negation, which concerns semantic or logical status, and *metalinguistic* negation, which concerns assertability. Crucially, descriptive negation is an operator over propositions, but metalinguistic negation relates to utterances.

Horn justifies the shift from semantic to pragmatic issues by presenting an array of data showing that the "presupposition-cancelling" variety of negative category mistakes and empty subjects — illustrated in (25) — is just one case of a more general phenomenon that is not restricted to the domain of semantics. For example, what is rejected in (29)

(29) Chris didn't manage to solve *some* of the problems, he managed to solve *all* of them.

is not a conventional implicature, but the *conversational* implicature that leads the typical listener to infer that "some" implies "not all."

Other conversational implicatures are handled along similar lines. For example, Grice (1975, p. 73) suggested that a reviewer might choose to write (30a) in order to imply that writing (30b) would have left out crucial information, such as a hideous defect in Miss X's performance.

(30) a. Miss X produced a series of sounds which corresponded closely with the score of 'Home Sweet Home.'
b. Miss X sang 'Home Sweet Home.'

According to Grice, the implication that the performance was terrible arises because the speaker has violated the submaxim of Brevity. Horn notes that the conversational implicature here can be rejected by negation just as in (29):

(31) Miss X didn't produce a series of sounds which corresponded closely with the score of 'Home Sweet Home,' dammit, she *sang* 'Home Sweet Home,' and a lovely rendition it was, too!

Implicatures associated with the order of conjunction (i.e. the correspondence of order to a time sequence or to importance) can be cancelled, as well:

(32) a. They didn't have a baby and get married, they got married and had a baby.
b. Mozart's sonatas weren't for violin and piano, they were for piano and violin.

Other examples show that the phenomenon is still more general, permitting the rejection even of phonetic, morphological, and stylistic aspects of utterances, or the focus or connotation implicated by a particular utterance.

(33) a. He didn't call the [pólis], he called the [polís].
b. I didn't manage to trap two mongeese, I managed to trap two mongooses.

(34) a. Now Cindy, dear, Grandma would like you to remember that you're a young lady: Phydeaux didn't shit the rug, he soiled the carpet.
b. It's not stewed bunny, dear, it's civet de lapin.

(35) a. I'm not his daughter, he's my father.
b. Ben Ward is not a black Police Commissioner but a Police Commissioner who is black.

Metalinguistic negation, then, constitutes "a formally negative utterance which is used to object to a previous utterance on any grounds whatever" (p. 374).[5] Unifying all these examples is a typical prosodic

[5]Horn's later analysis appears to weaken this statement somewhat; he isolates a class of implicata that cannot be cancelled by negation.

contour — "contrastive intonation with a final rise within the negative clause" — together with a continuation in which the faulty implicatum, whether lexical, morphological, or phonetic, is rectified. Horn strengthens his case for this analysis by showing that other logical operators have metalinguistic interpretations, as in:

(36) a. I can only very briefly set forth my own views, or rather my general attitudes. (From Sapir, *Language*)
b. If you're thirsty, there's some beer in the fridge.

and that his analysis also applies to scalar implicatures, where metalinguistic negation is used "for disconnecting the implicated upper bound of relatively weak scalar predicates" (p. 382):

(37) a. Around here, we don't *like* coffee, we *love* it.
b. That wasn't a bad year, it was a *horrible* year.

Finally, Horn describes a set of diagnostics for distinguishing descriptive from metalinguistic negation, including the inability of incorporated negation to license a metalinguistic reading, the ability of metalinguistic negation to permit positive polarity items, and the "archetypal" *not X but Y* frame for metalinguistic negation.

(38) a. The king of France is not happy.
b. The king of France is unhappy.

(39) a. Chlamydia is not "sometimes" misdiagnosed, it is frequently misdiagnosed.
b. #Chlamydia is not ever misdiagnosed, it is frequently misdiagnosed.

(40) a. It isn't hot, but scalding.
b. Negation is ambiguous not semantically but pragmatically.

In positing a distinction between the logical and the implicated aspects of an utterance, Horn's analysis is very much in the same spirit as that of Karttunen and Peters. However, Horn argues that unlike their account, his metalinguistic analysis is capable of handling not only objections to propositional content and conventional implicatures, but also objections based on improper grammar, choice of register, phonetics, and so forth. A similar point holds for analyses in which one reading of negation is taken to be a general operator more or less equivalent to "it is not true that S." Horn comments,

> Metalinguistic negation, as we have seen, is used to deny or object to any aspect of a previous utterance — from the conventional or conversational implicata that may be associated with it to its syntactic, morphological, or phonetic form. There can be no justification for inserting an operator TRUE into the logical form for a certain subclass of marked negative sentences, in order for negation to be able to focus on it, if metalinguistic negation does not in principle have to do with truth conditions. (p. 414)

Horn acknowledges that establishing a dichotomy between descriptive and metalinguistic negation opens up an enormous formal problem:

> One important question which I did not, and will not, directly address here is just how metalinguistic negation is to be represented within a formal theory of natural language discourse . . . We must be content for now with the negative fact extracted from this chapter: some

> instances of negation in natural language are not formally representable in an interpreted propositional language. (p. 444)

Since category mistakes and metalinguistic negation are so closely related, this last comment of Horn's might lead a cynic to consign them to a black hole (the one containing all those "pragmatic" phenomena that are deemed to be beyond formal treatment, at least for the next few centuries). In Section 3.4 I will argue for a more optimistic view, to be developed in the remainder of the chapter. First, however, I will briefly review the treatment of category mistakes in generative linguistic theory.

3.3 Selection Restrictions

3.3.1 Selection restrictions as lexical features

The phenomena discussed in the previous section appeared in the guise of "selection restrictions" in (Katz and Fodor, 1964). Katz and Fodor proposed a decompositional theory of word meaning in which lexical entries specified the features applicable to a particular lexical item. The classic example is the noun *bachelor*, decomposed into four lexical entries of the following form:

(41) a. (Human), (Male), [who has never married]
b. (Human), (Male), [young knight serving under another knight's standard]
c. (Human), (Male), [who has the first or lowest academic degree]
d. (Animal), (Male), [young fur seal when without a mate during the breeding time]

Features not encoded directly in the lexical entry were also taken to be part of a word's meaning — for example, something HUMAN is also ANIMATE — though specific mechanisms for accomplishing this (inheritance, redundancy rules) are not relevant here. (The items in parentheses are semantic markers, and elements in square brackets are "distinguishers." Semantic markers are intended to be primary, with distinguishers not constituting components of meaning *per se*; however, see (Fodor, 1977, Chapter 5) for a critical discussion of this view. Kastovsky (1980) comments that Fodor and Katz's distinction between markers and distinguishers "was completely rejected later as untenable both on theoretical and empirical grounds" (p. 86).)

For words that denote predicates, Katz and Fodor proposed that the arguments in their lexical entries (properly, variables in argument position) be annotated with selection restrictions — that is, specifications identifying the necessary and sufficient condition for a semantically acceptable argument. Such conditions were represented as Boolean functions of semantic markers; for example, (42) gives their selection restrictions on the arguments for the verb *hit* when used as in "The man hits the ground with a rock."

(42) a. [SUBJECT] (Human) $\vee$ (Higher Animal)
b. [OBJECT] (Physical Object)
c. [INSTRUMENTAL] (Physical Object)

Given a characterization of arguments as in (41) and of selection restrictions as in (42), the predicates and arguments were combined straightforwardly: the cross-product of all possible combinations would be taken over the senses associated with the component words in an expression, and selection restrictions would rule out inappropriate readings from the resulting set. For example, suppose *bachelor* had the four readings given above, *hit* had the reading in (42) plus one other (e.g. the reading in "the rock hit the ground with

a thud"), and *baseball* had two senses (e.g. as a physical object and a game). Under these circumstances, "The bachelor hit the baseball" would begin with sixteen ($= 4 \times 2 \times 2$) possible readings. However, some of those readings would be ruled out: assuming that seals are not higher animals, the selection restriction on the subject in (42) eliminates all readings in which that sense of *hit* takes (41d) as its subject. Similarly, the interpretive procedure would discard all readings construing *baseball* as a game.

The theory presented by Katz and Fodor (1964) has a number of important features. First, it accounts for the semantically anomalous character of selectional violations (i.e. category mistakes): if *no* sense of an argument meets the conditions on an argument of the predicate, then the set of readings for an expression will be empty. Second, it shows how selectional properties can be used in a positive way to constrain ambiguity, since the cross-product of all possible readings is reduced by those that are selectionally inappropriate. And third, it accounts, via decompositional lexical semantics, for the intuition that certain lexical combinations are redundant or tautologous — for example, the lexical decomposition of *bachelor* specifies that bachelors are unmarried, so the modifier in *unmarried bachelor* adds no new information when the two lexical entries are combined.

3.3.2 Selection restrictions as syntactic features

Chomsky (1965) adopted a theory of selectional restrictions that was in many ways similar to Katz and Fodor's, but he located selectional features in the syntactic rather than the semantic-interpretive component. Selectional constraints applied to lexical insertion, preventing anomalous predicate-argument combinations from being inserted into deep structures. The presence of two contemporary formal mechanisms, one syntactic and the other semantic, was complicated even further by the fact that Chomsky's syntactic view of selectional phenomena represented a shift from his semantic characterization in *Syntactic Structures* (Chomsky, 1957). Fodor (1977, p. 97) comments that "the treatment of selection at this stage was schizophrenic."

Chomsky did not express a particularly strong commitment to the syntactic treatment of selection restrictions; for example, he wrote:

> Selectional rules play a rather marginal role in the grammar, although the features that they deal with may be involved in many purely syntactic processes . . . One might propose, therefore, that selectional rules be dropped from the syntax and that their function be taken over by the semantic component. Such a change would do little violence to the structure of grammar as described earlier. (Chomsky, 1965, p. 153)

However, he did argue forcefully that selectional *features* such as [Human] and [Abstract] could not be excluded from the syntactic component, under the assumption that expressions like "the book who you read" are deviant on syntactic grounds.

As for the status of selectional rules themselves, Chomsky's arguments are neither particularly vehement nor particularly convincing. The main point he makes seems to be a response to the observation that semantic but not syntactic anomaly is acceptable in embedded contexts. He comments that although placing selection in the syntactic component would require an account for such cases as (43a), some such explanation would be necessary anyway, in order to account for the acceptability of embedded subcategorization violations as in (43b):

(43) a. It is nonsense to speak of frightening sincerity.
b. It is nonsense to speak of elapsing a book.

Thus at least one motivation for removing selectional rules from the syntactic component is not as strong as it might be.

(Although (43b) certainly requires an explanation, it seems to me that embedded contexts are irrelevant here. Rather the phenomenon seems to concern the question of how verb meanings and syntactic frames productively interact — as, for example, in:

(44) a. Gepetto danced Pinnochio (across the table).
b. The shortstop looked the runner back to second base.

See (Fisher, Gleitman, and Gleitman, 1991) for discussion, and cf. Grimshaw's (1993) suggestion that diathesis alternations are the result of core verb meanings interacting with clausal structure.)

3.3.3 Selection restrictions as semantic constraints

The discussion of selection restrictions in syntactic terms seems for the most part to have ended in the 1970s, owing at least in part to the influence of critiques by McCawley (1968) from the perspective of generative semantics. McCawley provides convincing evidence that selection involves at least semantics, contra Chomsky (1965); and contra Katz and Fodor (1964) he finds severe problems with their analysis of selectional phenomena based on a lexical decomposition theory of meaning.

As a first point against the lexical-features analysis of selectional phenomena, McCawley observes that selection restrictions must be construed as applying not to lexical items, but to entire constituents. He points out that the anomaly of (45a) cannot be attributed to the head noun of the subject, since (45b) is perfectly fine.

(45) a. My buxom neighbor is the father of two.
b. My neighbor is the father of two.

Furthermore, McCawley argues, selection restrictions can take into account any piece of semantic information about a lexical item, and not just some restricted set — he points out that many words have extremely specific selectional restrictions, such as *devein* (a shrimp or prawn), *diagonalize* (a matrix), and *benign* (a tumor, in medical usage). In addition, he claims that *only* semantic properties can serve as selectional features, so that apparent cases of selection for syntactic features (e.g. mass vs. count nouns) are really cases of selection for the semantic features with which they are correlated (e.g. whether or not the items referred to are individuated). Notice, for example, that there are no verbs that select just syntactically feminine subjects (e.g. in English, women and ships), but certainly some that select for the semantic feature FEMALE.[6]

Taken together, these points show that accounting for selection restrictions syntactically would duplicate much of the work already being done by the semantic component. This is nicely illustrated by the following example (given in (Fodor, 1977, p. 98)):

(46) a. This corpse admires sincerity.
b. This dead man admires sincerity.
c. This man that I proved that John was mistaken in believing to be alive admires sincerity.

Fodor notes that "the fact that the subject phrase of [46c] refers to a dead man is determined by the meanings of *prove*, *mistaken*, *believe*, and *alive* and by the way in which these words are combined . . . in other words, by the SEMANTIC content of the whole noun phrase."

[6] For further debate, see (Katz, 1970; McCawley, 1971).

Regarding Katz and Fodor's use of selection restrictions in disambiguation, McCawley argues convincingly that selectionally anomalous sentences must be assigned semantic representations, since, as noted earlier, anomalous constituents can be part of non-anomalous sentences. As a result, Katz and Fodor's mechanism of identifying semantic anomaly with an empty set of readings is untenable.

McCawley also shows that the disambiguation mechanism proposed by Katz and Fodor leads to unacceptable results. He notes that if *king* has just two readings, one related to monarchy and the other to chess, then when (47) is interpreted,

(47) It is nonsense to speak of a king as made of plastic.

the monarch reading will be ruled out, leaving only the unintended reading, "It is nonsense to speak of a chess piece king as made of plastic."

In general, McCawley takes issue with disambiguation based only on a privileged set of semantic markers identified with the meanings of words. He points out that the word *priest* has the semantic feature MALE associated with it not as an element of meaning, but as a fact about the current state of the world, since otherwise discussions about allowing women to become priests would necessarily concern sex-change surgery (cf. allowing aunts to become uncles). On Katz and Fodor's theory, MALE therefore cannot be a semantic marker for the word, and thus can play no role in disambiguation. However, this is contrary to the intuition that

(48) The landlord knocked the priest up.

is easily disambiguated, rejecting the "caused to become pregnant" reading in favor of the reading "The landlord awakened the priest by knocking on his door."[7]

3.3.4 Selection restrictions and inference

McCawley's position on (48) raises a difficult issue concerning the status of selection restrictions in linguistic theory. Although he does not say so explicitly, it appears that he takes the disambiguation of this sentence to be an application of selection restrictions — in this case, a selection restriction based on a property that he has just identified as "based on factual information rather than purely on meaning" (McCawley, 1968, p. 130).

Kastovsky (1980), discussing McCawley's position, argues that this is indeed the case. He attributes to McCawley the following contrast:

(49) a. My arm is bleeding.
b. The arm of the statue is bleeding.

and asserts that, similar to the *priest* example, the oddness of (49b) is based not meaning-related properties but on "extralinguistic probabilities." In particular, he says, "[–BLOOD] is not part of the inherent feature specification of *statue*" (p. 74).

On the one hand, it seems to me that Kastovsky is not giving enough credit to the compositionality of meaning and to inter-feature relationships: [–ANIMATE] is uncontroversially an "inherent" feature of

[7] This is a British usage: "If you knock someone up, you knock on the door of their bedroom or of their house during the night in order to wake them" (Sinclair (ed.), 1987). The point still stands for American English, of course: with the feature MALE unavailable as a selectional feature of *priest*, it would not be possible to account for the intuition that speakers of American English find the sentence anomalous.

statues, and inferring the absence of blood from the absence of animacy falls within the scope of whatever mechanism permits the inference of ANIMATE from HUMAN, since *blood* is definitionally associated with vertebrates and invertebrates, hence living things, at least according to the *American Heritage Dictionary*. Thus an appropriate compositional interpretation of *the arm of the statue* would lead to a true selectional violation in (49b). On the other hand, as Kastovsky points out, this may not be the desired result. A "minor extralinguistic miracle" could render the sentence perfectly acceptable; indeed, people have been known to establish shrines around statues of religious figures because they believed those statues, though inanimate, were literally bleeding.

Kastovsky's position leads to a fairly complicated state of affairs: how is one to determine which features of meaning are "inherent" and which are not? For example, according to Kastovsky, (45b) stands in contrast to (45a) because the feature MALE is inherently a part of *father* and presumably also because the feature FEMALE is inherently a part of *buxom*. Yet one could imagine talk-show host Geraldo Rivera introducing a guest in the following way:

(50) Now introducing John Smith, looking lovely after his breast-augmentation surgery. This buxom father of two makes his living in Las Vegas as a female impersonator...

Similarly, if another "minor extralinguistic miracle" of the medical variety permitted Jane Jones to produce sperm cells, Geraldo would almost certainly seek to have her on the show, and he might justifiably make the following introduction:

(51) Jane Jones, former supermodel, is now a buxom father of two.

Examples like these would seem to suggest that notion of "inherent" features is not clear cut — and in fact most predicates are likely to be even less clear than those involving features MALE and FEMALE.

The problem is not restricted to cases involving miracles of one kind or another. Drange (1966) illustrates the difficulty in distinguishing ordinary false sentences from semantic anomalies with the following series of sentences:

(52) a. Englishmen like coffee better than tea.
b. Squirrels like coffee better than tea.
c. Protozoa like coffee better than tea.
d. Bacteria like coffee better than tea.
e. Milkweed plants like coffee better than tea.
f. Stones like coffee better than tea.
g. Electrons like coffee better than tea.
h. Quadratic equations like coffee better than tea.

He comments,

> [Perhaps] this difference is not so much a difference in kind as a difference in degree. Sentences (a)–(h) seem to be arranged in a graded series in such a way that it is not at all clear where a line is to be drawn to distinguish the "factually incorrect" sentences from the "semantically incorrect" sentences (or the false from the meaningless). (p. 16)

Ultimately Drange argues that the difference is in fact a matter of kind — i.e. that one can draw a clear line distinguishing false expressions from type crossings in (52). However, his criterion for where to draw the line appeals to a notion he calls "unthinkability," for which I find his arguments unconvincing.

Drange acknowledges that even if (52) does illustrate a difference in kind, it may *also* illustrate a difference in degree. He comments,

> There are many properties which Englishmen share with things which like coffee better than tea that are not possessed by quadratic equations, such as being a physical object, being alive, having organs of taste, and so on. On the other hand, there are very few properties which quadratic equations share with things which like coffee better than tea, and it seems that Englishmen have all of them also. It is in this sense that the difference between (a) and (h) might be said to be one of degree. (p. 17)

There are two things to note about this statement. The first is that property comparisons of the kind Drange is discussing need not involve inherent or definitional properties — having organs of taste is neither a definitional aspect of human beings, animals, etc. nor a necessary component of the verb *likes*. Indeed, when Drange includes tastebuds in the discussion he is acting on a fact about the world, namely that a preference for coffee over tea will be determined on the basis of taste and not, say, color.

The second thing to note is that the kind of property comparison Drange suggests, taken in its simplest form, will be unilluminating. He points out that any concept possesses an infinite number of irrelevant property descriptions of the following kind:

(53) a. The property of not being composed of exactly one stone.
b. The property of not being composed of exactly two stones.
c. The property of not being composed of exactly three stones.

Since any two concepts share an arbitrarily large number of properties, the notion of "fewer" or "more" properties in common will not suffice.

The discussion thus far seems to suggest that selection restrictions may involve inference about factual knowledge at least to some extent. A more forceful argument, in favor of viewing selection restrictions as unequivocally inferential, is made by Johnson-Laird (1983), who writes that "the notion that it is possible to formulate exhaustive and definitive selectional restrictions on the different senses of words turns out to be a fiction" (p. 234). He supports this view using the following example:

(54) Alcock and Brown were the first to fly X from the USA to Ireland.

(55) a. Alcock and Brown were the first to fly an aeroplane from the USA to Ireland.
b. Alcock and Brown were the first to fly a bicycle from the USA to Ireland.
c. Alcock and Brown were the first to fly the Atlantic from the USA to Ireland.

Johnson-Laird notes that it is necessary to capture at least the three senses of *fly* illustrated here, which suggest argument selectional restrictions along the following lines:

(56) a. x (human, animal, or machine) controls the path through the air of y (vehicle)
b. x (human) takes y (physical object) in an aircraft.
c. x (physical object) travels in the air over y (geographical region)

The problem, he points out, is that these restrictions fail to constrain the arguments properly, permitting a sentence like (57a) to have interpretation (57b):

(57) a. I saw the Azores, flying the Atlantic.
b. I saw the Azores as they were flying over the Atlantic.

It seems clear that ruling out flying islands by appealing to a more specific selectional constraint in (56c) will be difficult. And, more to the point, doing so leads to a converse problem when the context supports such an interpretation, as, for example, in a science fiction story about how the earth explodes.

The alternative that Johnson-Laird proposes relies on inferential processes involving factual knowledge. He argues that the crucial question in interpreting (57a) is whether it is possible for the Azores to travel by air over the Atlantic ocean, a question that "can only be decided by making an implicit inference based on general knowledge." That is, the question "hinges on questions of fact, as well as on a knowledge of the meanings of words" (p. 235). The information Johnson-Laird has in mind is the following:

(58) a. An island is a land mass entirely surrounded by water.
b. Land masses are parts of the earth's surface that are fixed relative to other such parts (barring earthquakes).
c. An ocean is a body of salt water that covers a large and relatively fixed part of the earth's surface.
d. If x is a fixed part of y then x travels when y travels, but x does not travel with respect to y.
e. The Azores are islands in the Atlantic Ocean.

Notice that information about word meaning is not formally distinguished from factual knowledge; also note how nullifying the second piece of information — e.g. in the context of a science fiction story — would change the interpretation of (57b).

Although strictly semantic information is not formally distinguished from general world knowledge, one could argue that some selectional restrictions make reference only to senses of expressions — for example, the subject of *love* as human or animal. Johnson-Laird expresses skepticism about this claim, noting that "in a context where the dish ran away with the spoon, there may be nothing anomalous about a chair's falling in love with a table" (p. 236). He suggests that what appear to be selectional constraints based on semantic knowledge are, in fact, just cases of the more general process, with some inferences conventionalized because of their frequency and predictability. The inferential mechanism is the same, he concludes, whether the premises involved concern linguistic or factual knowledge.

3.4 Summary and Prospects

3.4.1 Properties of selectional constraints

On the basis of the discussion in the preceding two sections, I will characterize selectional constraints by enumerating a number of properties for which I find the evidence convincing, and noting issues that seem to me to remain unresolved.

The following properties seem adequately demonstrated:

1. Selectional constraints hold of constituents and not simply lexical items. This is demonstrated by the contrast between *neighbor* and *buxom neighbor*, only the second of which would violate a requirement for the feature MALE.

2. Selectional constraints are not strictly syntactic. This is demonstrated, e.g., by the existence of predicates selecting for semantic features (e.g. semantic rather than syntactic gender), and of verbs like *devein* that refer to semantic classes (e.g. shrimp and prawns).

3. Selectional constraints are not restricted to a small, privileged feature vocabulary. Examples like *devein* and *diagonalize* demonstrate that verbs can select for small, specific semantic classes rather than large, abstract ones.

4. Selectional violations cannot be interpreted as meaningless in the strict sense of a total absence of readings; first, because even selectionally deviant sentences generate entailments, and second, because distinct anomalous constituents in embedded (especially report and belief) contexts lead to distinct interpretations of the full sentence.

5. Selectional constraints are applied in a positive way to assist lexical disambiguation. This seems to be convincingly demonstrated by Katz and Fodor, despite the flaws in their particular disambiguation mechanism as noted by McCawley.

Other properties of selectional constraints seem unresolved. In particular:

1. Are selectional constraints restricted to "semantic" properties? Katz and Fodor's treatment, Chomsky's syntactic variant, the arguments of Drange and Kastovsky, and the bulk of the philosophical literature reviewed by Horn all would seem to agree that selectional constraints have to do with (at most) elements of meaning, and not general knowledge. McCawley seems to admit factual knowledge at least into the discussion of disambiguation (e.g. the probable but not definitional maleness of priests), and perhaps also into selectional constraints (this is Kastovsky's interpretation, at least); Horn groups negative category mistakes in with a pragmatic class of metalinguistic negations, but doing so does not preclude the possibility that the categories involved in selectional phenomena may refer only to elements of meaning. Johnson-Laird argues that selectional constraints should be viewed as part of a general inferential framework, and that putatively semantic selectional properties cannot ultimately be distinguished from factual constraints.

2. Are selectional constraints categorical or graded? Few of the authors explicitly discuss whether or not the selectional status of an argument is an all-or-nothing matter. When selection is taken to refer exclusively to semantic properties, the constraint usually takes the form of a categorical, necessary and sufficient condition and no other alternative is considered. When factual properties are discussed, there are some passing references to the probabilistic nature of such facts (e.g. Kastovsky's mention of "extralinguistic probabilities"), but it is not clear whether uncertainty, inferential support, or other aspects of extralinguistic processing are intended to result in partial or graded satisfaction of constraints. Of the authors considered here, only Drange explicitly discusses the apparent gradedness of anomaly judgements. Whether or not the satisfaction of selectional constraints is a matter of degree, the positive application of selectional constraints in interpretation (items 4–5, above) suggests that arguments may interact with selectional constraints in a more flexible way than is usually supposed.

3.4.2 A dilemma

The unresolved issues present a dilemma to someone interested in elaborating an empirically adequate theory of selectional constraints. Suppose, for the sake of argument, that one were to accept the view that selectional

constraints represent necessary and sufficient conditions, phrased only within a semantic vocabulary; that is, that they concern only senses of expressions and not any accompanying factual knowledge. Such an approach seems clearly to be equivalent to adopting a decompositional, definitional theory of description for mental categories. On such a theory, a category is completely defined in terms of (a Boolean combination of) features such as ANIMAL, FLIES, HAS WINGS, and so forth — the only difference is that in the case of selectional constraints the definition refers not to an independent category label (e.g. BIRD) but to applicability conditions for the argument position of a predicate (e.g. `x` in `blue(x)`).

Definitional theories, however, suffer from some well known problems. Armstrong *et al.* (1983, p. 268) report:

> [The] definitional theory is difficult to work out in the required detail. No one has succeeded in finding the supposed simplest categories (the features). It rarely seems to be the case that all and only the class members can be picked out in terms of sufficient lists of conjectured elemental categories. And eliminating some of the apparently necessary properties (e.g., deleting *feathers*, *flies*, and *eggs* so as to include the down-covered baby male ostriches among the birds) seems not to affect category membership. Generally speaking, it is widely agreed today in philosophy, linguistics, and psychology, that the definitional program for everyday lexical categories has been defeated — at least in its pristine form.

Johnson-Laird's (1983, chapter 10) discussion of selection restrictions, described briefly above, is of course also a critique of the definitional approach.

It is possible to adopt a decompositional theory without requiring necessary and sufficient conditions; this is the usual interpretation of prototype theory (Rosch et al., 1976). However, Armstrong *et al.* (1983) point out that, to the extent that a prototype theory makes use of features, it will have many of the same problems as a definitional theory: "it is not notably easier to find the prototypic features of a concept than to find the necessary and sufficient ones" (p. 272). Furthermore, in contrast to one of the main advantages of definitional theories, adopting a theory of prototypes makes difficult (Armstrong *et al.*: "altogether hopeless") a compositional account of phrase and sentence meanings (Osherson and Smith, 1981).[8]

A review of the literature here would constitute too much of a side-trip, unfortunately; for a start see (Fodor et al., 1980; Smith and Medin, 1981; Armstrong, Gleitman, and Gleitman, 1983; Smith and Osherson, 1988),. The main point here is that even if constraints are not viewed as necessary and sufficient, the identification of an adequate, exhaustive set of primitive selectional (more generally, semantic) features seems on empirical grounds to be difficult if not impossible to sustain.

Now consider the alternative view, the position that selectional constraints are connected not with semantic primitives and meaning but with inferences and factual knowledge. This is the line suggested by Johnson-Laird, and it seems consistent with Horn's classification of negative category mistakes within a much broader set of pragmatic issues. A theory of selectional constraints on this view becomes equally problematic, albeit for entirely different reasons. Where for a definitional theory the problem is not being able to find a complete and adequate set of features within the confines of the semantic representation, for a pragmatic or inferential theory the problem is that anything goes: it will be necessary to represent and make inferences about not only word meaning proper but also other facts ranging from social mores to naive

[8] However, see (Kamp and Partee, in progress) for a defense of the prototype theory. They ascribe many of Osherson and Smith's criticisms concerning compositionality not to prototype theory *per se* but to the choice of fuzzy logic as a supporting mechanism, and they propose an alternative formulation of prototype theory having a different probabilistic/semantic substrate that appears to resolve many of the problems.

physics. To say that such a theory is not within our current reach would be an exercise in understatement: researchers in artificial intelligence sometimes call problems like this "AI-complete," signifying that a solution would be tantamount to solving the AI problem itself.

3.4.3 A proposal

The dilemma I have just described has to do with the vocabulary in which selectional constraints are expressed: at one extreme the vocabulary consists of a relatively small set of semantic primitives, and at the other extreme selectional constraints can bring in practically the entire representational arsenal of human reasoning. I see no way to reconcile the two positions — if we unlock the door to conceptual rather than than strictly semantic representations and processes, I see no principled way to avoid opening it to its widest extent. Instead, I am going to propose a solution that avoids this difficult issue, and at the same time remains well motivated and empirically sound.

Since there is no in-between position, I will accept a dichotomy between the semantic and inferential viewpoints and focus on the latter. Although this may seem dismissive on the semantic side, I prefer to think of the distinction as analogous to the dichotomy Horn draws between descriptive and metalinguistic negation: a category mistake may be taken to have a simple truth value in descriptive terms, but the interesting part of the action is outside the formal semantics.[9]

The proposal has two components.

1. The first component is a taxonomic representation of noun concepts of the kind discussed in Section 2.4.1, in which the central relationships are hyponymy (IS-A) and synonymy. Although an IS-A taxonomy can be interpreted in many different ways, recall that I grounded the formalization in a relationship that might be called "plausible entailment": following (Lyons, 1961), the entailment relation is based on the ordinary judgements of the language user, such that "one sentence implies another if in saying the one we are prepared to say the other" (Sparck Jones, 1964, p. 54). The synonymy of two words — properly, word senses — hinges on the existence of representative sentences in which the words can be substituted while still preserving exactly the same set of plausible entailments.[10]

   This choice of representation will allow me to bypass the most serious unresolved issue, namely the problem of how to coherently discuss selectional constraints in terms of general inference without first understanding how to represent the knowledge behind such inferences. The notions of "features" or "properties" have no place at all in the taxonomy: if two words are companions in a synonym set, then by definition there is some set of representative sentences in which they are mutually substitutable according to ordinary judgements. However, and this is the crucial point, the mechanism for *making* those judgements is entirely irrelevant. What matters is that the taxonomy exist, not that the criteria for individual classifications be fully specified. General inferential mechanisms may ultimately account for the taxonomy, but the theory of selectional constraints I propose need not explain how.

[9] There may also be an analogy here to Kamp and Partee's (in progress) proposal for integrating prototype theory with truth-theoretic semantics: there, real-valued characteristic functions represent "constraints on the possible completions of a two-valued partial model" (p. 27). In what follows, selection will constitute a probabilistic relationship over a space of conceptual classes; perhaps Kamp and Partee's proposal could be extended so that this probabilistic framework also serves as a constraint on relationships within a truth-theoretic model.

[10] I have been deliberately vague about what I mean by "representative." The best characterization I can come up with is another appeal to ordinary judgements: a context is representative if an ordinary speaker would agree that the usage is not particularly creative or unusual. Although this is unsatisfying, it is no worse than the admonition to interpret a sentence "literally" in order to identify the reading on which it is anomalous (Drange, 1966, p. 12).

Simply stated, then, my hypothesis is that a conceptual taxonomy of this kind implicitly encodes (most of) the inferences needed to account for selectional constraints. This leaves open the possibility that, although the taxonomy is a useful construct for present purposes (i.e. until a correct theory of inference is available), people really do apply selectional constraints in the form of inferences computed using factual knowledge. A stronger version of the hypothesis is that a taxonomy of this kind is psychologically real, and that the "true" theory of selectional constraints really does involve stored conceptual knowledge rather than on-line inferences.

Notice that this strategy accommodates Drange's objection to property comparisons (Drange, 1966, p. 17). He quite correctly observes that any two concepts will share an infinite number of properties in common. A definitional theory provides one way out, by restricting the relevant properties to an exhaustive, finite set of features. The strategy adopted here provides a different way out, via the generally useful tactic of reducing infinites to a finite set of equivalence classes.[11] Notice also that although the taxonomy represents not "linguistic" but "factual" relations between words, it does so in a minimal way, adhering to what Miller calls the "standard lexicographic line" in distinguishing between lexical concepts and general knowledge (see footnote 14 in Chapter 2). As I have noted, that such a minimal extension will suffice is admittedly no more than a hypothesis, but I hope the remainder of this chapter will show it to be a plausible one.

2. The second component of my proposal concerns the formalization of selectional relationships within the vocabulary of the taxonomy I have just described. One straightforward approach would be simply to reformulate Katz and Fodor's theory using the conceptual classes themselves as features. This option would improve on their account of selection restrictions by greatly expanding the base of primitives, in accord with McCawley's arguments (see properties 2 and 3, above). However it must be rejected, I think, since like the semantic theory it fails to account for the meaningful interpretation of selectionally deviant utterances in embedded contexts, the partial meanings ascribed to selectional violations even in matrix utterances, and the subtle effects of selectional constraints on lexical disambiguation (see properties 4–5).

   The alternative, which I will elaborate in the next section, moves away from idea of restrictions, and toward a characterization of selectional phenomena in terms of preferred association.[12] The way in which an argument satisfies and fails to satisfy the preferences of a predicate will account for how it is interpreted, whether or not the combined expression constitutes what would traditionally be called a selectional violation. In order to accomplish this, some formal means of representing preferences will be necessary. Having already moved away from traditional analyses by eliminating semantic features, I will diverge still further in the next section by formalizing the preference relationship probabilistically using the tools of information theory.

[11] To state this more precisely, let W be the set of word forms. On the theory proposed here, the set of properties may be infinite, but the set of equivalence classes of properties is indexed by $\mathcal{P}(W)$ and therefore finite. (See Table 2.3.)

[12] This may parallel the relationship between selectional restrictions and "lexical solidarities" discussed by Kastovsky (1980).

3.5 Selection as Information

3.5.1 Intuitions

The alternative view of selectional constraints I am proposing can be phrased as follows: rather than restrictions or hard constraints on applicability, a predicate preferentially associates with certain kinds of arguments, and these preferences constitute the effect that the predicate has on what appears in an argument position. For example, the predicate `blue` does not *restrict* itself to arguments having a tangible surface — the sky is blue, and so is ocean water even deep below any apparent surface — but its arguments are still far from arbitrary. The effect of the predicate is that its arguments tend to be physical entities and to have surfaces. Similarly, the verb *admire*, interpreted in the particular sense "to have a high opinion of," has an effect on what appears as its subject; these tend to be physical, animate, human, capable of the higher psychological functions, and so forth, though it may well be that no Boolean combination of these properties is both necessary and sufficient. In some cases the effect a predicate has on its argument is quite strong: one is unlikely to find the (numerical) predicate `even` applied to anything but positive integers, though zero and the negative integers are also fairly likely. In other cases — e.g. the predicate `smooth` — the effect is less dramatic.

Expressions that fail to observe preferences are nonetheless interpreted in accordance with them. For example, if someone told you in all seriousness that

(59) Milkweed plants like coffee better than tea,

you might think they were uttering something absurd, but you could legitimately expect *them* to believe that milkweed plants are capable of feeling pleasure, expressing a preference, or some other property typically expected of the subjects of *like*. Similarly, as McCawley (1968) points out,

(60) My aunt is a bachelor.

may raise an eyebrow, but it will nonetheless typically be interpreted as meaning that the aunt is unmarried. Thus the violation of one expectation does not necessarily imply that others are not met. Incidentally, if the speaker were just beginning to learn English, one might plausibly adjust one's belief according to a different preference if the context supported it; e.g., concluding that *aunt* was mistakenly used in place of *uncle*. This is the kind of context dependence stressed by Johnson-Laird (1983).

Katz and Fodor's (1964) use of selection restrictions in disambiguation can be recast as an inferential process based on preferences. In the straightforward cases, its behavior is very much the same; for example, in

(61) John hit the baseball.

the reading of *baseball* as a physical object will be a better match for the preferences of *hit* than its reading as a kind of game. In addition, though, the inferential view makes sense of some of the problematic cases that McCawley noted, such as (60), above, and

(62) It is nonsense to speak of a king as made of plastic.

In the latter case, what is crucial is that the applicability of the predicate MADE-OF-PLASTIC to *king* is ruled out not locally by compositional semantics — leading erroneously to the interpretation "It is nonsense to

speak of a chess piece king as made of plastic" — but as part of a global inferential process, which takes into account the embedding of the proposition within the matrix predication "It is nonsense that S."

I am not going to attempt a formal characterization of the inferential process I have been discussing, though see (Johnson-Laird, 1983, chapter 11) for a theory of comprehension that is a good match for the views presented here. The point I wish to stress is that the issues under consideration here are prerequisite to *any* theory of interpretation. All the theories of category mistakes reviewed by Horn (1989) presuppose that the question "can predicate P apply to argument x?" has a yes-or-no answer, and any theory consistent with the discussion of examples (59) through (62) will require some way of answering the more general question "what could P(x) mean?"

3.5.2 Formalization

Prior and posterior distributions

Rephrasing the problem in terms of preferences naturally suggests a probabilistic treatment. As a prerequisite, however, it is important to be clear about two distinctions concerning the status of probabilities.

The first distinction concerns the "logical" versus the "empirical" view of probability. The theoretical treatment here is based on the former conception — something like the probabilities one would find by peering into the head of the language user. This stands in contrast to empirical view, in which probabilities are defined in terms of what one would observe as the experimental sample grew infinitely large.[13] Second, one must always distinguish probabilities from statistical probability *estimates* based on observed samples. If a fair coin comes up heads six times and tails four times, the probability p(heads) is still exactly $\frac{1}{2}$ even though the usual probability *estimate* in this case would be $\frac{3}{5}$.

Formally, let P be a random variable ranging over the set $\{p_1, \ldots, p_m\}$ of predicates under consideration, and let C be a random variable ranging over the set $\{c_1, \ldots, c_k\}$ of classes in the taxonomy, with C be related to P by a particular predicate-argument relationship, such as subject-verb, verb-object, or adjective-noun.[14] Given this probabilistic framework, the intuitive notion of "preference" can now be phrased more precisely as the following question: what effect does the choice of a particular predicate p_i have on the distribution of C?

Figure 3.1 illustrates how this might work for a particular verb, *grow*, with respect to classes of direct objects. The top of the figure represents what the distribution of argument classes might be regardless of the particular predicate. As the figure shows, some classes are *a priori* simply more likely to be referred to in direct object position, and some less likely; for example, absent any other information, animals might be more likely to be mentioned in direct object position than legumes. However, given the particular verb *grow*, this distribution changes to the one shown at the bottom of the figure: some classes (e.g. animals) become much less likely, and others (e.g. legumes) become much more likely.

It is this relationship, the change between the *prior* distribution, $\mathrm{p}(c)$, and the *posterior* distribution, $\mathrm{p}(c|p_i)$, that constitutes selectional preference. On this account, the features or properties that govern selectional constraints remain entirely hidden. Selectional relationships are characterized entirely by the

[13] There is a long history of debate concerning the inductive (or logical) view of probability as distinguished from empirical probability. The first chapter of (Bulmer, 1967) contains one extremely brief but useful introduction to the distinction. See also (Bar-Hillel, 1964, chapters 15 and 16) for discussion specifically with regard to information theory.

[14] I will only be considering predicates corresponding to surface syntactic relationships, but this is easily generalized. I will also consider only one argument of a predicate at a time, which will lead to empirical difficulty in some cases. For example "The dinosaur devoured the village" and "The mouse devoured the village" will differ in selectional status with regard to the direct object, owing to the nature of the subject.

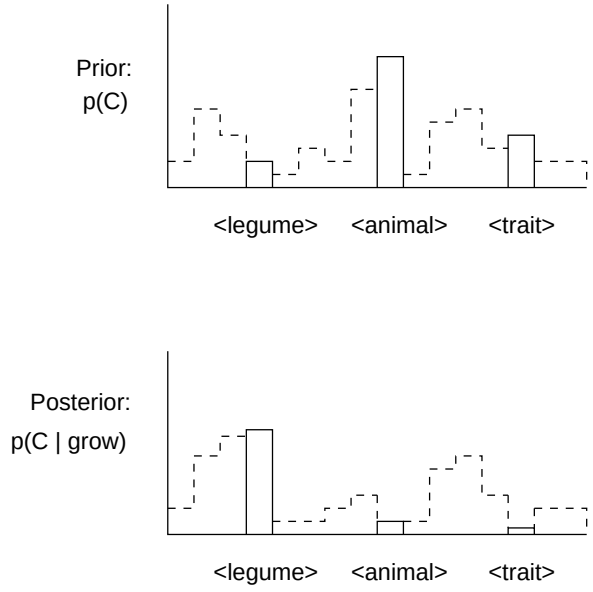


Figure 3.1: Example of a prior distribution and a posterior distribution

probabilistic relationship between a predicate and the classes of its arguments. Since classes in the taxonomy are defined in terms of "plausible entailment," general inferential processes play a role in this characterization. However, this role is an indirect one, and there is no need to presuppose that those processes are represented or even representable.

Constraints and Information

The probabilistic characterization of selectional relationships just described relies on the notion of a "difference" between two probability distributions. This idea can be formalized using the tools of information theory (Shannon and Weaver, 1949), providing not only a precise definition but also an illuminating interpretation of what it means for a predicate to constrain an argument "weakly" or "strongly."

Recall from the discussion in Section 2.2 that in information theory the *entropy* of a random variable is a measure of how uncertain the outcome is, on average. For example, if X represents the possible outcome of a fair coin flip, then its entropy $\mathrm{H}(X)$ will be high. If the coin is unfair — having, say, a 90% to 10% bias in favor of coming up heads — then the entropy of X will be quite low. To take a more relevant example, suppose W ranges over nouns in English, and that it corresponds to the next word given the introduction

(63) The cook basted the ____

The entropy of W in this case will be relatively low, since it is overwhelmingly likely that the next word will be one of a small set of words such as *turkey* or *roast*. On the other hand, if W is introduced by

(64) The cook enjoyed the ____

then its entropy will be much higher, since any of an enormous number of completions is reasonably likely — the chef might enjoy a book, the opera, or the company of the butler. (The use of entropy to measure the predictiveness of contexts like these is discussed in (Treisman, 1965; van Rooij and Plomp, 1991).)

In information theory, entropy or uncertainty is generally identified with quantity of information. To understand why this correspondence makes sense, consider how the informational state changes when an actual event occurs, if you already knew the underlying probability distribution. In the case of the heavily biased coin, actual flips tell you little more than you already knew: it will tend to come up heads, which is *unsurprising*, given the distribution, and therefore conveys very little *information*. On the other hand, the fairer the coin, the less information you begin with: since you don't have any idea what to expect, every flip of a completely fair coin will be maximally informative. Examples (63) and (64) are analogous.

Formally, entropy (information) is defined as:

$$\mathrm{H}(X) \quad = \quad \sum_x \mathrm{p}(x) \log \frac{1}{\mathrm{p}(x)}, \tag{3.1}$$

and the quantity of information obtained by observing a single x is equated with $\log \frac{1}{\mathrm{p}(x)}$. Notice that, consistent with the intuitive description just given, the more surprising (less probable) something is, the more informative it will be.[15] With that in mind, entropy can be seen as informativeness "on average": $\mathrm{H}(X)$ is a weighted average of information taken over all possible values for X. Since the logarithm is conventionally taken to the base 2, the standard unit of information is the *bit*, short for "binary digit." (I will not elaborate on the reasoning behind the particular mathematical form in (3.1); see (Khinchin, 1957, pages 9–12) for a nice exposition of this point and (Cover and Thomas, 1991) for a very readable introduction to information theory as a whole.)

Relative entropy is an information-theoretic measure of how two probability distributions differ, which is precisely the question under consideration here. Given two probability distributions p and q, their relative entropy is defined as

$$\mathrm{D}(p \parallel q) \quad = \quad \sum_x p(x) \log \frac{p(x)}{q(x)}. \tag{3.2}$$

Like entropy, the relative entropy is a weighted average; moreover, rewriting equation (3.2) as

$$\mathrm{D}(p \parallel q) \quad = \quad \sum_x p(x) [\log \frac{1}{q(x)} - \log \frac{1}{p(x)}] \tag{3.3}$$

makes it clear that what is being averaged is the difference at each point between information according to distribution q and information according to distribution p. A useful interpretation of the definition comes from thinking of probability distributions as models: relative entropy can be interpreted as the cost, measured in bits of information, of using q as a model when the true distribution is p.

Notice that under this interpretation, it would not make sense for relative entropy to ever be negative: any model q that is not exactly correct should incur a positive cost relative to the perfect model of p, namely p itself. And this is, in fact, the case: an important theorem of information theory is that $\mathrm{D}(p \parallel q)$ is *always* greater than or equal to zero, and equal to zero if and only if $p = q$ (Cover and Thomas, 1991, p. 26).

This fact can seem counterintuitive, since equation (3.2) shows that relative entropy is the sum of many terms of the form $[p(x) \log \frac{p(x)}{q(x)}]$, each of which may be positive (when $p(x) > q(x)$), negative (when $q(x) > p(x)$), or zero. Although it might seem as if the positives and negatives could balance out even when p and q are different, this turns out not to be the case. Consider two simple examples, supposing

[15] Since $\log \frac{1}{\mathrm{p}(x)} = -\log \mathrm{p}(x)$, this is the same as the definition given in Chapter 2.

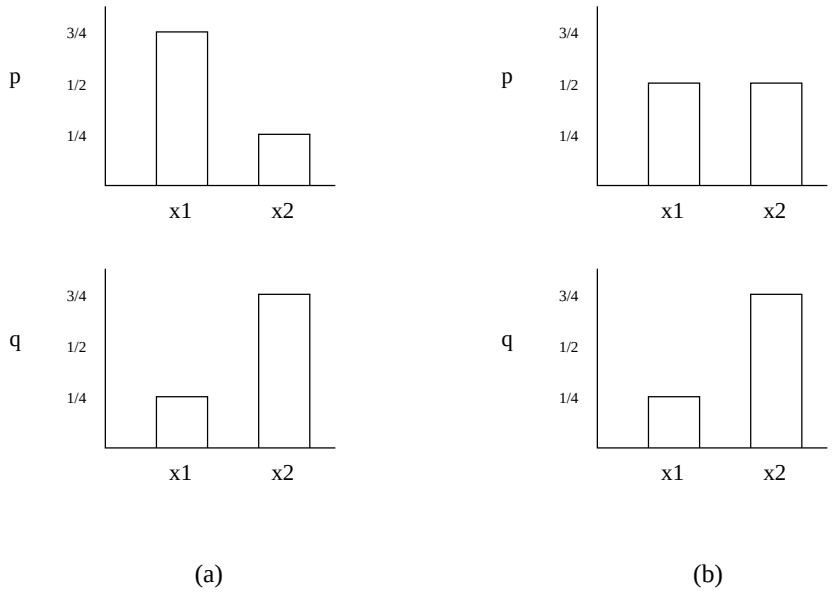


Figure 3.2: Simple distributions to illustrate relative entropy

that x can take on only two values, x_1 or x_2. First, consider the case where $p(x_1) = \frac{3}{4}$, $p(x_2) = \frac{1}{4}$, and where $q(x_1) = \frac{1}{4}$, $q(x_2) = \frac{3}{4}$ (see Figure 3.2(a)). Here, the probability of x_1 is lower in q than in p, and the probability of x_2 is higher by exactly the same amount. Although the two might therefore seem to balance each other out, the relative entropy between the two distributions is positive; specifically,

$$\begin{aligned} \mathrm{D}(p \parallel q) &= \sum_x p(x) \log \frac{p(x)}{q(x)} \\ &= \frac{3}{4}\log 3 + \frac{1}{4}\log\frac{1}{3} \\ &= \frac{3}{4}\log 3 - \frac{1}{4}\log 3 \\ &= \frac{1}{2}\log 3. \end{aligned}$$

As a second example, suppose that x_1 and x_2 have the same probability in p, and differ from that value in the model q by equal but opposite amounts (see Figure 3.2(b)). In this case, although both x_1 and x_2 have equal weight in the "true" distribution, and their values in q change in equal and opposite directions, the relative entropy nonetheless turns out to be positive:

$$\begin{aligned} \mathrm{D}(p \parallel q) &= \frac{1}{2}\log 2 + \frac{1}{2}\log\frac{2}{3} \\ &= \frac{1}{2}(\log 2 + \log 2 - \log 3) \\ &= \frac{1}{2}(\log 4 - \log 3) \\ &= \frac{1}{2}\log\frac{4}{3}. \end{aligned}$$

Working through further examples like these, it should become clear that, despite one's intuitions, it is impossible to "trick" the definition into balancing out for non-identical distributions!

Relative entropy can be applied straightforwardly to the probabilistic treatment of selectional constraints. The prior distribution of classes, $\mathrm{p}(c)$, represents an "uninformed" model of what the distribution of arguments looks like, one that does not take the predicate into account at all. The posterior, $\mathrm{p}(c|p_i)$, is the true distribution of argument classes for a particular predicate p_i. So, treating the former as q and the latter as p in equation (3.2), the difference between the two distributions is quantified as:

$$\mathrm{D}(\mathrm{p}(c|p_i) \parallel \mathrm{p}(c)) \quad = \quad \sum_c \mathrm{p}(c|p_i) \log \frac{\mathrm{p}(c|p_i)}{\mathrm{p}(c)}. \tag{3.4}$$

In information-theoretic terms, equation (3.4) measures the information provided about a random variable (C, the class of the argument) by an event ($P = p_i$, i.e. observing the predicate). Smyth and Goodman (1992) discuss why this measure, which they call the j-measure, is the appropriate one to use for that purpose.

At this point, I will try to heed the admonition of Bar-Hillel (1964, chapter 15) against confusing various ideas of what information is. Although Shannon and Weaver (1949) explicitly tried to avoid interpreting "information" in terms of semantic content, Bar-Hillel notes that their use of the term has often not been carefully distinguished from the notion of meaning, with undesirable results. In the present work, I have followed Bar-Hillel in adopting the logical view of probabilities for theoretical purposes, though for computational purposes the theory is of necessity implemented using statistical probability estimates. I think this succeeds in addressing one of Bar-Hillel's concerns. The other major concern is to adequately distinguish between the *quantity* of information and semantic *content*. To be perfectly clear, I am identifying the *selectional preference* of a predicate with the overall difference (or "change") between the prior distribution of argument classes and the posterior distribution conditioned on the predicate. Although quantitative, the definition preserves content — this is what I hope to have conveyed visually in Figure 3.1. Equation (3.4) concerns quantity rather than content: it encapsulates the difference between the distributions as a scalar, measured in bits of information, that I will call *selectional preference strength*.

Selectional preference strength is very much like the idea of selectional range — intuitively, some predicates are selectionally more restrictive than others. However, since I have dispensed with explicit features, it is no longer possible to identify selectional range extensionally as the set of arguments for which a Boolean combination of features is true. Instead, on the model just proposed, the intuition that a predicate's selectional constraints can be narrow or wide, strong or weak, has an information-theoretic interpretation.

Consider a predicate that strongly constrains the nature of its arguments, i.e. one that would intuitively be said to have a narrow selectional range. In this case, the posterior distribution — the distribution of argument classes conditioned on the predicate — will be very different from the prior distribution, with those classes that satisfy the predicate's preferences increasing their share and classes that fail to satisfy it decreasing in probability. As a result, the relative entropy will be high and the predicate will have a high selectional preference strength. For a different predicate that places weaker constraints on its arguments, the overall difference between the two distributions will not be as great. As a result, the selectional preference strength will be low.

Perhaps most interesting, in this model the selectional preference strength of a predicate is not just a number, but a number with a precisely specified meaning. As discussed above, relative entropy is measured in bits of information, and can be interpreted as the cost of assuming that the distribution is q when the real distribution is p. When p and q are assigned as in equation (3.4), this translates into the cost of assuming the distribution is $\mathrm{p}(c)$ when it is really $\mathrm{p}(c|p_i)$ — that is, the cost of not taking the predicate into account. Therefore in a very direct way, the selectional preference strength of a predicate can be understood as the amount of information that it carries about its argument. I will explore this interpretation further in Chapter 4.

3.6 Predicted Behavior

Since selectional preference is now in some sense a relationship between a predicate and the entire conceptual space of arguments, it is no longer clear what it would mean for a particular argument to satisfy a selectional preference, or to violate one. This is not surprising given the inferential view that being taken here, since, as Johnson-Laird (1983) argues, the ability of an argument to appear with a predicate is less a yes-or-no decision and more a function of how easily the predication can be accommodated given information about word meanings and context. In addition, it is consistent with the intuition, noted by Drange (1966), that judgements of selectional fit are a matter of degree rather than categorical. In accord with this view, rather than attempting to define the notion of an argument satisfying a predicate, I will consider in the probabilistic setting what the consequences are for interpretation when a predicate is applied to an argument.

I will assume, or rather continue assuming, that in the lexicon each noun is mapped to a set of concepts in the taxonomy, corresponding to its different senses. For example, the noun *baseball* might be mapped to two concepts, one which is a hyponym of the concept ⟨ball⟩ (in the sense of a "round object that is hit or thrown or kicked in games"), and the other of which is a hyponym of the concept ⟨field game⟩.[16] A noun will be said to "belong to" any class in the taxonomy having one of its concepts as a hyponym, directly or indirectly. Thus *baseball* belongs to not only the class ⟨ball⟩ but also to others such as ⟨game equipment⟩, ⟨artifact⟩, and ⟨entity⟩, by virtue of its first sense; and by virtue of its second sense it belongs to ⟨outdoor game⟩, ⟨sport⟩, and ⟨human activity⟩, among others. In addition, I will assume that a compositional procedure exists for mapping noun phrase arguments to sets of concepts in the taxonomy — such a procedure would, for example, yield different mappings for the arguments *my arm* and *the arm of the statue*.

Unlike arguments, predicates will be treated simply as symbols. One might argue that they, too, fit within a taxonomy, but this is not a point I will pursue further.

I will take for granted the existence of the prior and posterior probability distributions described above, reminding the reader that these are theoretical (abstract, logical) probabilities rather than empirical probability estimates. (I might also point out that the prior distribution is the same for all predicates, though for the sake of discussion it is useful to talk about each predicate as having both a prior and a posterior.) Given these distributions, the selectional preference of the predicate — that is, the difference between the two distributions — can be seen as consisting of a data point for each class in the taxonomy, corresponding to each term of the sum in equation (3.4). (The strength of preference is therefore merely the result of adding all these points together.) For example, using Figure 3.1 as a model for the distributions for *grow*, three of those data points are:

(65) a. $\mathrm{p}(\langle \texttt{legume} \rangle | \textit{grow}) \log \frac{\mathrm{p}(\langle \texttt{legume} \rangle | \textit{grow})}{\mathrm{p}(\langle \texttt{legume} \rangle)}$

b. $\mathrm{p}(\langle \texttt{animal} \rangle | \textit{grow}) \log \frac{\mathrm{p}(\langle \texttt{animal} \rangle | \textit{grow})}{\mathrm{p}(\langle \texttt{animal} \rangle)}$

c. $\mathrm{p}(\langle \texttt{trait} \rangle | \textit{grow}) \log \frac{\mathrm{p}(\langle \texttt{trait} \rangle | \textit{grow})}{\mathrm{p}(\langle \texttt{trait} \rangle)}$

Given a particular predicate and its distributions, I will call the data point corresponding to a class c the selectional behavior of the class with respect to the predicate. This forms the basis for a measure that I will

[16] For the remainder of this discussion I am adopting labels from the WordNet taxonomy, as indicated by angle brackets.

call *selectional association*:

$$\mathrm{A}(p_i, c) \quad = \quad \frac{1}{\eta_i}\, \mathrm{p}(c|p_i)\, \log \frac{\mathrm{p}(c|p_i)}{\mathrm{p}(c)}. \tag{3.5}$$

The selectional association of a predicate for an argument class is simply the "data point" just under discussion, with the additional complication of a divisor that is constant given the particular predicate. The divisor will simply be the selectional preference strength of the predicate, as defined in equation (3.4); that is,

$$\eta_i \quad = \quad \sum_c \mathrm{p}(c|p_i) \log \frac{\mathrm{p}(c|p_i)}{\mathrm{p}(c)}. \tag{3.6}$$

It is included in the definition in order to obtain a measure of predicate-argument association on a scale that is in some sense independent of how strongly or weakly the predicate selects overall.[17]

When a noun appears as the argument of a predicate, the relevant question is now clear: what is the selectional behavior of the classes to which the noun belongs? For some of those classes, the predicate determines little or no change between the prior and posterior distribution; in those cases, $\mathrm{A}(p_i, c)$ will be relatively small.[18] For some classes, the posterior probability leaps up compared to the prior probability; in those cases, the value of $\mathrm{A}(p_i, c)$ will be positive and the class can be said to be "selected for" (to a greater or lesser extent). Finally, a third set of classes will exhibit a marked drop between the prior and the posterior probability; in those cases the class can be described as "selected against." It should be noted that since changes are weighted by the posterior probability, classes that are selected against will have less of an overall influence on selectional preference strength than classes that are selected for.

What is particularly important here is that the selectional behavior must be considered not for any particular class to which an argument belongs, but for each of those classes. This provides the basis for some important observations about selectional behavior in cases that would traditionally be construed in terms of an argument "satisfying" or "violating" the selection restriction of a predicate.

First, consider what happens when the argument satisfies the selectional restriction for a predicate with intuitively strong selectional constraints (e.g., *eat turkey*). In such a case, there will clearly be some class or set of classes that the predicate "selects for," in the probabilistic sense just described. The good fit of the argument to the predicate can be determined by its membership in a class for which the selectional association is high.

Second, consider what happens when the argument satisfies the selection restriction for a predicate with intuitively weak selectional constraints (e.g. *enjoy movie*). Here, the probabilistic behavior is similar to the previous case, except that the selected-for classes are not marked as clearly by dramatic shifts between the prior and posterior distribution. In this case, the noun is a member of at least one of the selected-for classes, but the value of selectional association is comparatively low, though positive, for even the strongest of them.

Third, consider what happens when the argument violates the selection restriction of a predicate. If the argument is selectionally inappropriate in the *intended* sense, but appropriate in another, *unintended* sense, then it is the latter sense that will emerge on the basis of selectional association. For example, consider the interpretation of

(66) The music is brown.

[17] Note that including η_i does not result in $0 \leq \mathrm{A}(p_i, c) \leq 1$, since the contribution of a class to selectional preference strength may be either positive or negative. Clearly, however, $\sum_c \mathrm{A}(p_i, c) = 1$ for all p_i.

[18] This is not perfectly accurate, since the change is weighted by the conditional probability $\mathrm{p}(c|p_i)$. For classes that are strongly predicted given the verb, even relatively small differences between the prior and posterior may be magnified significantly.

On the expected interpretation of *music* as a member of classes ⟨sound⟩, ⟨sense experience⟩, and so on, this is a standard selectional violation, a fact that will be reflected in zero or negative selectional association between *brown* and those classes. However, *music* can also be interpreted as a physical object, e.g. a ⟨creation⟩ or ⟨artifact⟩. *These* classes fit the selectional restriction of the predicate, and thus they will emerge as the classes for which selectional association is positive, with the magnitude being greater or less according to how strong or weak the constraints are. As a result, the selectionally appropriate interpretation of *music* as an argument for *brown* — its selectional profile, if you will — emerges from the predicate's association with selected-for classes to which the word belongs. (Cf. example (12) in Section 3.1.)

At first glance, there would seem to be a fairly direct analogy between the process just described and the Katz-Fodor account of disambiguation based on selection restrictions. However, it must be stressed that the present account locates selectional preference within the inferential component rather than within local and strictly semantic interpretation. Thus other factors, such as the contextual appropriateness of alternative senses of the argument, may certainly play a role. If the discourse or situational context preferentially supports the interpretation of *music* in its physical-object sense — for example, if my friend has found the libretto of *The Mikado* on my shelf and is now looking for the score — I can utter (66) and it will be selectionally perfectly appropriate in its intended sense. And if the context makes unavailable the alternative reading of *music* — for example, if I am listening to classical music on the radio — then (66) will sound anomalous.[19]

In addition, considering another example makes it clear that what is going on is closer to the flexible accommodation that McCawley describes in interpreting

(67) My aunt is a bachelor.

than to a disambiguation procedure based on the strict matching of selectional features. Although the word *aunt* does not belong to the class ⟨male⟩, it certainly belongs to ⟨person⟩, with which *bachelor* will also be associated. Thus the selectional profile of *aunt* in the context of *bachelor* may support the successful interpretation of (67), albeit not as strongly as *uncle* would have. Note, too, that adopting the account of selectional preference proposed here does not preclude the encoding of necessary (or even definitional) descriptions of predicates when they are available; thus (67) might be (to some extent) acceptable at the selectional or inferential level, but at some other level ruled anomalous.

3.7 Empirical Behavior

3.7.1 Computational apparatus

I have explored the empirical behavior of the model by means of a computational implementation. The implementation is fairly faithful to the theoretical model, with the exception of several simplifications that are for the most part unproblematic.

The most significant modification, of course, is in the nature of the probabilities used. It is possible to construct the theoretical foundations of a model according to the logical view of probabilities, but in practical terms the probabilities of the model must be estimated statistically on the basis of some source of evidence. As discussed in Section 2.2, I have adopted the statistical technique of maximum likelihood estimation, not

[19] I have not worked out the interaction of selectional preference with context. The most straightforward approach, I think, would be to assume that there is some finite set Γ of relevant contextual features and to condition all the probabilities on $\gamma \in \Gamma$; e.g. to calculate selectional preference strength as $\mathrm{D}(\mathrm{p}(c|p_i,\gamma) \mid\mid \mathrm{p}(c|\gamma))$.

only for its simplicity, but also to avoid presupposing a solution to the problem of estimation from sparse data; however, that concern is not relevant here. I am convinced that the behavior of this model is not particularly sensitive to the estimation technique used, at least at a general level, since in earlier versions of this work I computed probabilities using the Good-Turing estimate (Good, 1953) and obtained entirely comparable results. (See Appendix A.)

The selectional relationship I have explored most thoroughly is the one that holds between verbs and their direct objects. In the work I will describe here, estimates of verb-object co-occurrence probability were arrived at by constructing a large combined sample from the following sources:

1. Associated Press (AP) news stories (1989)
2. The Brown corpus (Francis and Kučera, 1982), specifically the parsed version of the corpus appearing in the Penn Treebank (Marcus, Santorini, and Marcinkiewicz, 1993)
3. The *Wall Street Journal* (1988-89), which can also be found in parsed form in the Penn Treebank.
4. Transcribed parental speech from the Child Language Data Exchange (CHILDES) database (MacWhinney and Snow, 1985; Sokolov and Snow, to appear)
5. Verb-object norms collected from human subjects.[20]

In addition to verb-object relations, I have to some extent explored subject-verb relationships, using the Brown corpus and the Associated Press samples; adjective-noun combinations, using the Brown corpus and the *Wall Street Journal*; noun-noun modification, also using the Brown and *Wall Street Journal* corpora; and preposition-object relationships, using the Brown and CHILDES corpora.

Using samples of naturally occurring text as the basis for probability estimates required two additional simplifications. First, since I have no procedure for compositionally interpreting noun phrases, probabilities were estimated from a sample containing verbs together with the *head* of the direct object noun phrase. Although in principle this could lead to misleading results (you don't buy soldiers, although you might buy a *toy* soldier), modifiers rarely seem to have such a radical effect on the classes of the head noun. Second, since text corpora are not lexically disambiguated in advance, I have treated predicates as atomic symbols, which often conflates the selectional behavior of multiple senses. For example, the selectional behavior of the verb *play* is influenced in this model by the fact that it appears not only with direct objects like *baseball* and *game*, but also with *piano* and *violin*, and with *role* and *part*.

This second simplification is not as troublesome as it might first appear. When a verb has several very different senses, its distribution of argument classes tends to have distinct "clumps," that is, to appear multimodal. However, the existence of several strongly preferred categories is not necessarily a problem, since the interpretation of a single predicate-argument combination takes into account only those classes within which the argument appears. So, for example, the classes ⟨`musical instrument`⟩, ⟨`artifact`⟩, and so forth will have little bearing on the interpretation of *play baseball*, and the classes ⟨`sport`⟩, ⟨`activity`⟩, and so forth will have little bearing on the interpretation of *play piano*. Furthermore, the criteria for distinguishing verb senses are at present so poorly understood that avoiding terms like "homonymy," "polysemy," and so forth could be viewed as appropriately cautious rather than inappropriately simplistic.[21] Finally, assuming

[20] I am extremely grateful to Donald Hindle for making the AP data available to me, and to Annie Lederer for making available the verb-object norms. Notice that all subcorpora except the last one contain naturally occurring data.

[21] This may be a classic instance of the computer programmer's claim, "It's not a bug, it's a feature!" — but sometimes the programmer is correct.

that predicates were distinguished in advance would lead to a chicken-and-egg problem with respect to language acquisition — for example, see (Gropen, 1993) for an argument that selectional constraints are a necessary component in children's acquisition of multiple senses for a single verb. (Also see Chapter 4.)

I have adopted WordNet (Beckwith et al., 1991) as a computational model of the noun taxonomy, as discussed in Chapter 2. Although like any other dictionary WordNet has its idiosyncracies, in principle it has been constructed according to the theoretical taxonomic model argued for here. I believe this makes it unique among other dictionaries, certainly dictionaries that are available on line for computational purposes.

3.7.2 Traditional examples

I will organize this discussion around examples that have occurred over the course of the preceding sections.[22] To begin with, consider some of the classic examples of category mistakes, such as:

(68) a. The number two is blue.
b. Socrates is even.

As it happens, both *two* and *Socrates* are included in the WordNet taxonomy, despite its understandably limited coverage of numbers and proper names. Space precludes a detailed presentation of the entire selectional profile for any predicate-argument combination, but in general looking at the single class that *maximizes* selectional association will provide a good idea of how the model is behaving. This is what is shown in the following table:

(69)

Predicate	Argument	Maximum	Class
blue	two	-0.16	⟨measure⟩
even	two	3.99	⟨number⟩
blue	Socrates	2.66	⟨entity⟩
even	Socrates	0.03	⟨person⟩

Clearly *two* is an inappropriate argument for *blue*, since even the most strongly associated class to which it belongs is selected against (i.e. has a negative selectional association with the predicate). On the other hand, the application of *even* to *two* is fine. Conversely, Socrates is in selectional terms an appropriate argument for *blue*, by virtue of being a physical entity, though note the class ⟨person⟩ has a selectional association of -0.13 (not shown in the table). When *even* is applied to Socrates, the resulting selectional association is relatively indeterminate; I suspect that the marginally positive value results from noise in the sample from which probabilities were estimated.

To take two of Chomsky's best known examples,

(70) a. Colorless green ideas sleep furiously.
b. Sincerity may admire the boy.

the resulting selectional behavior is summarized as follows:

[22] In this section, I have multiplied all values of selectional association by 100 in order to avoid a needless proliferation of decimal places, and as a matter of notation I will represent the application of predicate P to argument x as Px. Probabilities for the adjective-noun and subject-verb relationships described in this section were estimated using the Brown corpus; for verb-object relationships I used the collection of samples described above.

(71)

Predicate	Argument	Maximum	Class
sleep	idea	-0.26	⟨psychological feature⟩
admire	sincerity	—	—

In both cases, the subject-verb relationship fails to yield a positive value for selectional association. The dashes arising from the predication *admire sincerity* indicate that the predicate never had as its subject any member of any class to which *sincerity* belongs; I regard such cases as indicating a selectionally inappropriate predication.

The behavior of the subject-verb model for

(72) a. Quadratic equations do not move in space.
b. Quadratic equations do not watch the Newmarket horse races.
c. Quadratic equations do not go to race meetings.

is captured in the following table:

(73)

Predicate	Argument	Maximum	Class
move	equation	-0.07	⟨deed⟩
watch	equation	—	—
go	equation	0.10	⟨communication⟩

As predicted, a description of equations as moving or watching will be selectionally unacceptable.[23] However, the behavior of the predication *go equation* leads to an interesting observation. Treating the subject of a verb as independent of the complement has led in this case to an unexpected result, for although equations cannot go to race meetings, one should not conclude that they cannot go. For example, it makes perfect sense to say that an equation should go at the top of the page. Thus, given only the limited information available to the model, its assignment of (weak) selectional plausibility to the predication in (72c) is not inappropriate.

Having considered some examples of selectional violations, I will turn to some of the other effects of selectional constraints. First, it must be reiterated that selectional constraints are viewed here as part of a more general inferential system, in contrast to the Katz-Fodor treatment of selection restrictions as semantic constraints on compositional interpretation. To take a concrete example of this, the verb-object combination *frighten sincerity* is ruled selectionally inappropriate by the implemented model (i.e. no class to which *sincerity* belongs has a positive selectional association with *frighten*), but this does not make

(74) a. It is nonsense to speak of frightening sincerity.

a selectional violation. I assume that the embedded predication would be evaluated in the context of the matrix clause, confirming — at least in part on the basis of the selectional relationship — that "nonsense" is an appropriate description.

A similar point holds true for a related example discussed earlier:

(75) It is nonsense to speak of a king as made of plastic.

[23] The "best" class is often not particularly meaningful if the selectional association is negative. Class ⟨deed⟩ contains words that denote completed actions; this includes *equation* in its sense as the act of making two things equal.

Although according to the implemented model the predication *plastic king* does indeed favor interpreting *king* as a chess piece (i.e., ⟨artifact⟩), the matrix clause in a sense reverses the interpretive process. In particular, an interpretive procedure, identifying conditions under which the sentence would be true, would presumably rule the "chess piece" interpretation out *because* it is selectionally appropriate and therefore *not* nonsense, at least on selectional criteria. On those same criteria,

(76) It is nonsense to speak of a king as being a monarch.

would be judged odd, since the acceptability of the predication in the embedded clause is inconsistent with what is being asserted about it in the matrix sentence.

Turning to other cases of disambiguation, consider the interpretations of *baseball* as both a physical object and as a game.

(77) a. I hit a baseball to John and he caught it.
b. I played some baseball yesterday afternoon.
c. I watched some baseball yesterday afternoon.

The verb-object relationships in these examples yield selectional relationships summarized in the following table:

(78)

Predicate	Argument	Maximum	Class
hit	baseball	4.48	⟨object⟩
play	baseball	2.73	⟨game⟩
watch	baseball	1.90	⟨diversion⟩

As a direct object for the verb *hit*, the interpretation of *baseball* as a physical object rather than a game emerges quite clearly. This becomes even more apparent in looking at the selectional profile for this predication; that is, selectional association of the verb with all the classes to which the argument belongs:

(79)

Value	Class containing *baseball*
4.48	⟨object⟩
4.27	⟨artifact⟩
2.25	⟨entity⟩
0.13	⟨ball⟩
0.11	⟨game equipment⟩
0.07	⟨equipment⟩
⋮	⋮
-0.00	⟨sport⟩
-0.00	⟨game⟩
-0.01	⟨diversion⟩
-0.01	⟨contest⟩
-0.01	⟨competition⟩
-0.27	⟨group action⟩
-0.34	⟨activity⟩
-0.85	⟨act⟩

When the verb is *play*, the profile is more or less inverted, with the classification of *baseball* as a game or competition emerging as the most strongly selected for by the verb. Interestingly, when the verb is *watch*, the most strongly associated class turns out to be ⟨diversion⟩:

(80)

Value	Class containing *baseball*
1.90	⟨diversion⟩
1.33	⟨object⟩
1.26	⟨artifact⟩
1.23	⟨equipment⟩
0.79	⟨competition⟩
0.66	⟨contest⟩
0.65	⟨act⟩
⋮	⋮

Interestingly, however, the selectional behavior of the predication, taken as a whole, shows quite a bit of ambiguity, since the "physical object" interpretation of *baseball* is also supported by strongly associated classes such as ⟨object⟩ and ⟨artifact⟩. This accords with my intuitions, since I can imagine taking my son to a park to watch baseball there, or (a few years from now) telling him to watch the baseball carefully as he's trying to catch it.[24]

The ambiguity of *score*, from the discussion of McCawley (1968) earlier, shows similar behavior in the implemented model.

(81) a. John has memorized the score of the Ninth Symphony.
b. The score of the Ninth Symphony is lying on the piano.

The word *score* is highly ambiguous: it has six senses in WordNet other than its musical interpretation, including among them a group of twenty things, a measure or abstraction ("the score is 2 to 1 in our favor"), a psychological feature (*score* as a kind of evaluation or assessment; also "facts about the actual situation," as in "He didn't know the score"), and an act or accomplishment ("He turns, shoots...score!"). Despite all those senses, the selectional profile of the predication *memorize score* clearly favors the intended interpretation:

(82)

Value	Class containing *score*
1.29	⟨musical composition⟩
1.04	⟨creation⟩
0.89	⟨music⟩
0.71	⟨art⟩
⋮	⋮
-0.13	⟨measure⟩
-0.15	⟨abstraction⟩
-0.15	⟨psychological feature⟩
-0.17	⟨group⟩
-0.19	⟨act⟩
-0.74	⟨entity⟩

[24] Clearly the determiner also plays an important role in disambiguation — "hit *the* baseball" vs. "play *some* baseball" — but the main point regarding selectional constraints stands regardless of whether additional evidence for resolving the ambiguity is also available.

In contrast, when *score* appears as the subject of *lie* in (81b), the selectional constraints favor what would appear to be entirely extraneous senses of the word.

(83)

Value	Class containing *score*
0.76	⟨psychological feature⟩
0.65	⟨abstraction⟩
⋮	⋮
0.14	⟨object⟩
0.07	⟨music⟩
0.07	⟨musical composition⟩
⋮	⋮

Looking at the sample of verb-object co-occurrences from which the model was constructed, this is easily explained. *Lie*, when used in subject position, very often concerns features of mental life (84) and abstract concepts (85):

(84) a. The only hope for West Berlin lies in a compromise which will bring down the wall and reunite the city.
b. The artistic interest, then, lies in what the encounter may be made to represent...
c. Regardless of where personal sympathies may lie as between the parties, failure to recognize these changed conditions would be to ignore the facts of life.

(85) a. The danger lay in the American delusion that nuclear deterrence was enough.
b. The cemetery slumbered just behind it, and the way lay through the village and close to the sea.
c. Although he still didn't speak to anyone, he grew fond of saying, " The future lies in Asia," when the opportunity arose...

The situation is further complicated by the "tell a falsehood" sense of *lie*, and the fact that people can appear as the subject for either sense.

All things considered, it is still worth noting that the sense of *score* as a physical object does nonetheless win out over its sense as a musical composition in (83). In addition, the story changes considerably if *score* is taken to be the *object* of either *lie* or *lay* (in some dialects the former can be used transitively to mean the latter):

(86)

Predicate	Argument	Maximum	Class
lie	score	4.06	⟨object⟩
lay	score	2.78	⟨object⟩

If John were to have laid the score of the Ninth Symphony on the piano, my example would have been much cleaner, but overall I believe the preceding discussion supports the point of view for which I have been arguing.

To take a final example, recall Drange's (1966) series of examples, which seem to suggest a graded rather than categorical distinction between empirical falsehoods and selectional violations.

(87) a. Englishmen like coffee better than tea.
b. Squirrels like coffee better than tea.
c. Protozoa like coffee better than tea.
d. Bacteria like coffee better than tea.
e. Milkweed plants like coffee better than tea.
f. Stones like coffee better than tea.
g. Electrons like coffee better than tea.
h. Quadratic equations like coffee better than tea.

(I recommend that before continuing, the reader decide for himself or herself the answer to the following question: if you were asked to divide these eight sentences into groups by drawing horizontal lines wherever you wanted to, where would you draw the lines?)

The results in the verb-object model are as follows:

(88)

Predicate	Argument	Maximum	Class
like	Englishman	5.35	⟨person⟩
like	squirrel	5.16	⟨life form⟩
like	protozoa	5.16	⟨life form⟩
like	bacteria	5.16	⟨life form⟩
like	milkweed	5.16	⟨life form⟩
like	stone	3.26	⟨entity⟩
like	electron	3.26	⟨entity⟩
like	equation	-0.30	⟨communication⟩

Looking at specific numbers, I am surprised by the strength of association between like and ⟨entity⟩ as its subject; at present I have no explanation for this. On the other hand, looking only at the groupings, the progression from people to life forms to entities to more abstract concepts strikes me as entirely reasonable, though given the choice I might take a more anthropomorphic view of squirrels and a less anthropomorphic view of milkweed plants. It seems to me that (88) is a thorough illustration of what I have been driving at: a definitional model of selectional constraints could draw only a single distinction in the above table, contrary to my intuition (and, I hope, the reader's as well); furthermore, to my knowledge no inferential model capable of making any more distinctions than that has been worked out in the necessary detail.

3.7.3 Argument plausibility

Another empirical test for the information-theoretic model of selectional constraints arises in the context of research into on-line processes during sentence comprehension. A great deal of recent work suggests that the plausibility of arguments plays a role in local syntactic disambiguation decisions: summarizing a review of relevant psycholinguistic studies, Ferstl (1993, p. 31) concludes that "selection restrictions seem to have an immediate effect in sentence processing."

These effects are demonstrated using stimuli like the following, taken from (Holmes, Stowe, and Cupples, 1989):

(89) a. The secretary read the article was already out of date.
b. The secretary read the fashion was already out of date.
(90) a. The scientist showed the sample was necessary for her project to succeed.
b. The scientist showed the travel was necessary for her project to succeed.
(91) a. The mayor recognized the author was worn out.
b. The mayor recognized the pocket was worn out.

Holmes *et al.* describe the distinctions in terms of (a) "plausible" versus (b) "implausible" objects, and in general psycholinguistic researchers seem to commit only to a notion that might be called "pragmatic plausibility," "argument typicality," or "local semantic fit," rather than selection restrictions as traditionally construed by linguists (see, e.g., (Boland et al., 1989; Tanenhaus, Garnsey, and Boland, 1991; Pearlmutter and MacDonald, 1993; MacDonald, in press; MacDonald, in revision; Tabossi et al., in press)).

In general, the psycholinguists determine the plausibility or typicality of verb-argument combinations by pre-testing. For example, Holmes *et al.* asked subjects to rate sentences like

(92) a. The tenant remembered the reply.
b. The tenant remembered the smoke.

on a 1-to-5 scale, and Tabossi *et al.* (in press) evaluated "agenthood" and "patienthood" by asking subjects to answer questions like

(93) a. How common is it for a reporter to interview someone?
b. How common is it for someone to interview a reporter?

on a scale from 1 to 7. These ratings can be used simply to confirm that the data in plausible-argument and implausible-argument conditions are in fact adequately distinguished, or they can be used as predictors of some aspect of on-line processing, such as reading time in a self-paced reading task.

Psycholinguistic studies of this kind are clearly relevant to the model of selectional constraints I have proposed. To the extent that plausibility or typicality ratings reflect selectional constraints rather than on-line inferential processes — and I think this may be to a great extent — judgements made by human subjects represent empirical data against which the model can be evaluated. Furthermore, to the extent that this model accurately reflects some aspect of human performance, it provides an important (and, I think, previously unavailable) methodological tool, not only for experimental design but also for the implementation of full-scale computational models of the psycholinguistic theories being proposed and debated. In this section I will explore the first of these issues, namely the evaluation of the model against human ratings; I consider the second in more detail toward the end of Chapter 4.

As a first step, I investigated the behavior of the implemented model using data from (Holmes, Stowe, and Cupples, 1989, Appendix 2) concerning verbs that have a bias in favor of taking NP complements. These consisted of sixteen pairs of sentences, three of which I have shown in examples (89)–(91). The verb-object combinations were constructed according to the experimenters' intuitions, and, as just mentioned, sentences like (92) were then rated for plausibility on a scale of 1 (low plausibility) to 5 (high plausibility) by human subjects. Holmes *et al.* report a mean rating of 4.5 for the sentences containing plausible objects, and a mean rating of 2.2 for implausible objects.

In Table 3.1 I show the sixteen verb-object combinations, together with the maximum value for selectional association in my implemented model and the class that achieved that maximum. As I noted earlier, selectional goodness of fit between a verb and an argument is in principle a function of the entire selectional

	Plausible			Implausible		
Verb	Object	Max	Class	Object	Max	Class
see	friend	4.42	⟨entity⟩	method	-0.02	⟨method⟩
read	article	6.26	⟨written comm.⟩	fashion	-0.00	⟨fashion⟩
find	label	0.06	⟨communication⟩	fever	2.35	⟨evidence⟩
hear	story	4.63	⟨communication⟩	issue	4.63	⟨communication⟩
write	letter	7.85	⟨writing⟩	market	0.31	⟨artifact⟩
urge	daughter	10.31	⟨agent⟩	contrast	0.31	⟨relation⟩
warn	driver	10.66	⟨agent⟩	engine	7.25	⟨entity⟩
judge	contest	1.58	⟨contest⟩	climate	0.22	⟨state⟩
teach	language	1.96	⟨cognition⟩	distance	1.76	⟨psych. feature⟩
show	sample	2.07	⟨psych. feature⟩	travel	0.92	⟨happening⟩
expect	visit	2.54	⟨act⟩	mouth	0.06	⟨reply⟩
answer	request	5.14	⟨speech act⟩	tragedy	3.70	⟨communication⟩
recognize	author	0.62	⟨agent⟩	pocket	-0.00	⟨concave shape⟩
repeat	comment	6.02	⟨social relation⟩	journal	6.02	⟨social relation⟩
understand	concept	4.13	⟨psych. feature⟩	session	2.69	⟨social relation⟩
remember	reply	0.00	⟨answer⟩	smoke	7.49	⟨entity⟩

Table 3.1: Selectional association for NP-bias verbs

profile, not just the single "best" class, but taking the maximum provides a useful and frequently accurate upper bound.

The values in the table are encouraging: mean maximum values for plausible and implausible objects are respectively 4.3 and 2.4, and the difference is significant ($t(15) = 1.9$, $p < .08$).[25] In general, the most strongly selected-for class of a plausible argument represents an entirely reasonable interpretation among a typically large number of classes — for example, the object in *read article* is interpreted as written communication rather than as a grammatical term and the object in *warn driver* is interpreted as a person rather than as a golf club.

Most of the mistakes in the table arise when an intuitively implausible object turns out to have some unanticipated, plausible use with respect to the verb. For example, *fever* is classified as a symptom and therefore as a form of evidence, and therefore plausible as the object of *find* — e.g.

(94) We brought Johnny to the doctor to find out why he's been so cranky, and the doctor found a fever together with a mild ear infection.

The word *distance*, as an object for *teach*, is being interpreted as a psychological feature, in the sense of emotional distance. Although one can construct sentences where this interpretation is not implausible,

(95) At the finishing school, the girls were taught not only the fine points of etiquette, but also distance and aloofness.

what is going on might be viewed as an undesirable overgeneralization. It arises because in WordNet 1.2, anything that is an instance of ⟨knowledge⟩ (e.g. *history*, *science*) is also an instance of ⟨psychological_feature⟩.

[25] The similarity of these numbers to the means obtained by Holmes *et al.* is purely accidental, of course; there is no relationship between their ratings scale and the scale used here.

	Plausible			Implausible		
Verb	Object	Max	Class	Object	Max	Class
say	phrase	-0.00	⟨phrase⟩	pencil	26.25	⟨entity⟩
know	teacher	10.31	⟨entity⟩	traffic	2.72	⟨group⟩
swear	oath	0.61	⟨curse⟩	exit	12.11	⟨entity⟩
argue	point	0.68	⟨content⟩	order	1.00	⟨act⟩
prove	theorem	1.47	⟨psych. feature⟩	battery	1.28	⟨abstraction⟩
forget	outcome	1.46	⟨psych. feature⟩	weekend	0.93	⟨time⟩
deny	charge	5.57	⟨speech act⟩	summer	-0.08	⟨time period⟩
claim	victory	1.99	⟨winning⟩	library	-0.00	⟨collection⟩
doubt	sincerity	1.76	⟨attribute⟩	champagne	0.03	⟨alcohol⟩
decide	match	2.96	⟨event⟩	award	0.19	⟨transferred property⟩
learn	truth	1.55	⟨cognition⟩	trial	1.55	⟨cognition⟩
realize	mistake	3.22	⟨psych. feature⟩	vehicle	0.03	⟨conveyance⟩
confess	fault	1.48	⟨act⟩	brake	-0.23	⟨artifact⟩
believe	witness	13.11	⟨life form⟩	journey	0.23	⟨change⟩
explain	decision	5.14	⟨psych. feature⟩	audience	0.90	⟨social relation⟩
discover	route	4.58	⟨object⟩	opera	4.58	⟨object⟩

Table 3.2: Selectional association for clausal-bias verbs

In other cases, the word *tragedy* is quite unexpectedly interpreted in its sense as a dramatic composition, and thus a form of communication, and, even more unusual, *smoke* is being interpreted as a physical object meaning *cigarette*. This leads to the counterintuitive ratings in the last line of the table, since in the sample from which probabilities were estimated, the direct objects of *remember* tend to be things having physical reality (especially people) rather than forms of communication.

As a second test, I used the data on clausal-bias verbs from (Holmes, Stowe, and Cupples, 1989) — that is, verbs that prefer a clausal rather than an NP complement. Table 3.2 shows the results: other than two whoppingly wrong decisions (for *say* and *swear*), the trend is clearly for the model to assign greater selectional plausibility to those verb-object combinations that were judged more plausible on intuitive grounds. On inspection of the corpora used to estimate probabilities, it becomes apparent why the model is so seriously misguided in those two cases. The AP sample of verb-object co-occurrences is, it appears, heavily contaminated by what is probably a systematic misanalysis for verbs like *say*, *report*, *swear*, *conclude*, etc., most likely when they appear inverted at the end of a clause, and perhaps also because they can introduce embedded clauses without an intervening complementizer. For example:

(96) a. “I’m innocent,” swore the prisoner as he was led to jail.
b. “AAAARRRRGGGH!” said Charlie Brown.

(97) a. The warden swore the prisoner was guilty as sin.
b. Snoopy says Charlie Brown needs to lighten up.

Of the twenty most frequent objects for *say* in that sample, thirteen are members of the class ⟨entity⟩; the distribution is enormously skewed toward the most frequent object, which is a special token denoting proper names (mapped to class ⟨person⟩ in the WordNet taxonomy, and therefore also a member of ⟨entity⟩). For *swear* fifteen of the top twenty objects are members of ⟨entity⟩. It also cannot help that the word *thing*, also a frequent object of *say*, is classified only as a physical entity in WordNet Version 1.2 (I believe

this is corrected in later versions); as a result the class ⟨entity⟩ gets erroneous support from examples like the following:

(98) a. The apostle Paul said the same thing...
b. One can make them say the same thing only by not listening to them very carefully and hearing only what one wants to hear.
c. How can you say such a thing?

Other than those most obvious problems, I find the results for this second group of examples analogous to the first experiment and therefore very encouraging.

A stronger result than what I have just described would be a correlation between plausibility ratings assigned by human subjects and the values assigned by the implemented model. This was not possible to test using the stimuli just described, since Holmes *et al.* reported only the average ratings for plausible and implausible combinations. However, other authors such as Tabossi *et al.* (in press) and Trueswell (1993) report the mean typicality rating for each of their test items, and I hope to use these data in future work.

3.8 Other Computational Approaches

Many computational approaches to selectional constraints have appeared in a form that is more or less similar to the view proposed in (Katz and Fodor, 1964): in implemented systems, something analogous to Boolean applicability conditions is often associated with each argument of a predicate.[26] For example, Schank (1986, p. 172) describes using "simple world knowledge rules" tied to conceptual rules, so that "the conceptual rule that actors can act would be modified by lists of what could do what according to semantic categories, such as 'animals can eat,' 'planes can fly,' and so on." Similarly, several of the natural language interfaces developed at BBN (e.g. see (Ayuso et al., 1989)) have used variants of the KL-ONE formalism to taxonomically represent world knowledge, implementing selectional constraints using that formalism's notion of "role restrictions" (Woods and Schmolze, 1991). It seems fair to say that a review of this approach to selectional constraints in a computational setting would in fact amount to a review of most of the natural language processing literature — this would no doubt raise a great many interesting issues (for example, cooperative responses when a selectional constraint is violated), but it is an enterprise I think it best to avoid here.

Of all the computational approaches with which I am familiar, Preference Semantics (Wilks, 1986; Wilks and Fass, 1992) is the one to which the present proposal seems most similar. Preference Semantics abandons the formalization of selectional constraints as restrictions — to use Wilks's (1986) term, "stipulations" — and instead interprets applicability conditions as preferences that can be satisfied or not satisfied and still yield some interpretation. Wilks (1986, p. 199) writes:

> It is very important to note that a preference is between alternatives. If the only structure derivable does *not* satisfy a declared preference, then it is accepted anyway.

A crucial component of preference semantics is the notion of "semantic density": the more preferences that are satisfied, the more preferred is the overall interpretation. For example, the sentence

[26] Often the practical approach to selectional constraints adopted in these systems is difficult to relate to formal semantic considerations of the kind discussed in Sections 3.2 and 3.3, though for an interesting exception see the computational approach to presupposition and entailment described in (Weischedel, 1986).

(99) The big policeman interrogates the crook.

might have two possible interpretations initially, one in which *crook* is interpreted to mean a criminal, and the other in which it denotes a shepherd's staff. In both cases, there are satisfied preferences between *big* and *policeman*, and between *policeman* and *interrogates*, but the latter interpretation has one fewer satisfied preference than the former since *interrogate* preferably describes something done by humans to humans.

It is important to note that these preferences are encoded in lexical entries that are essentially decompositional, expressed in a vocabulary of 80–100 primitive semantic units. For example, Wilks (1986) gives the following "semantic formula" for *interrogate*:

(100) ((MAN SUBJ) ((MAN OBJE)(TELL FORCE)))

This indicates a preference for humans as subject and object, and also indicates that the denoted action is one of forcing, in this case forcing to tell something. It is also important to note that the interpretation of sentences is accomplished by matching possible interpretation against abstract "semantic templates" that encode the possible structures that messages can take; for example, MAN FORCE MAN. Interpretations for which no matching template can be found are discarded — and as Wilks (1986, p. 197) points out, this commits Preference Semantics to the hypothesis that there is a "finite but useful inventory of bare templates adequate for the analysis of ordinary language; a list that can be interpreted as the messages that people want to convey at some fairly high level of generality."

The brief summary I have just given falls far short of an adequate description of Preference Semantics, but I hope it will suffice in order to identify the major points of similarity and dissimilarity with the proposal I have made in this chapter. The point of view I have adopted is to a very great extent consistent with the Preference Semantics enterprise; in particular, selectional constraints are discussed in terms of preference rather than restriction, and in Preference Semantics they are interpretable in quantitative terms via the notion of "semantic density" (though Wilks (1986) distances himself from a probabilistic viewpoint). At a more general level, the present proposal is in agreement with the stress in Preference Semantics on paying attention to "words of a normal vocabulary, and with many senses of them, rather than with single senses of simple object words and actions" (Wilks, 1986, p. 194).

There are also some important differences between the two approaches, most notably the question of lexical decomposition. Although I recognize that selectional constraints are only one part of a more general interpretive process, and therefore that some more elaborated representation of actions and participants will ultimately be necessary, I am distrustful of attempts to represent meaning exhaustively and decompositionally using a small(ish) set of primitives. Although I offer no alternative solution at present, I am encouraged in thinking that such a solution may be possible by the fact that the current model contains no explicit "enumeration" of selectional properties. For example, the selectional profile of *interrogate crook* is:

(101)

Value	Class containing *score*
9.38	⟨person⟩
9.35	⟨agent⟩
9.21	⟨life form⟩
6.33	⟨entity⟩
0.96	⟨wrongdoer⟩
0.95	⟨bad person⟩
0.44	⟨criminal⟩
0.09	⟨implement⟩
-0.02	⟨article of commerce⟩
-0.26	⟨artifact⟩
-0.37	⟨object⟩

which makes immediately apparent the correct reading of the direct object in (99). Although the disambiguation here is accomplished using a taxonomic representation of noun meaning, the preference of *interrogate* for its direct object argument is not an enumerated set of properties except in the sense that the prior and posterior probability distributions range over the entire conceptual space.

Turning to corpus-based models, much of the recent activity in statistical methods for natural language processing is related to the approach I have been pursuing, in the sense that any model involving lexical co-occurrence probabilities is, appropriately construed, a model of selectional constraints. A more clearly relevant subset is the set of probabilistic models involving word classes; these were the subject of the literature review in Sections 2.3 and 2.4. However, only a few recent proposals make explicit reference to selectional constraints. Grishman and Sterling (1992; 1993) have adopted a frequency-based approach: a relational triple (e.g. [*eat,subject,Fred*]) is ruled out on selectional grounds if it did not appear in the training data some minimal number of times. In their earlier paper, Grishman and Sterling used a manually-constructed noun classification hierarchy to generalize the selectional patterns in the training corpus; however in more recent work they have shifted to a smoothing technique based on a distributional measure of noun similarity (see discussion in Section 2.3). A similar approach is taken by Sekine *et al.* (1992), who cluster words on the basis of distributional similarity and use the clusters in identifying selectional patterns. A third approach of this kind is seen in the work of Velardi and colleagues (Velardi, 1991; Velardi, Pazienza, and Fasolo, 1991; Basili, Pazienza, and Velardi, 1991; Basili, Pazienza, and Velardi, 1992) — they, too, focus on the acquisition of relational triples, expressed using a relatively small set of semantic tags within a restricted domain.

I think the proposal I have made in this chapter differs from previous corpus-based approaches in a number of important ways. First, unlike most existing work on extracting selectional constraints from corpora, I have committed from the outset to working with unconstrained data rather than limited subdomains; as a result, the questions of how the proposal will "scale up" or how "transportable" it is are much less of a concern. Second, my emphasis has been on redefining the notion of selectional preference, not extracting a catalogue of selectional patterns from corpora. The latter, though an important problem, is more closely related to the "traditional" view of selectional constraints, since associating a set of classes with the argument of a predicate is equivalent to specifying a disjunctive (hence Boolean) applicability condition on that argument. A third closely related point is that I have attempted to make sure that the model of selectional constraints has a reasonably well-specified semantics, something that is not done in most computational proposals (though see (Velardi, 1991) for an interesting discussion on the relationship between corpora and various forms of

semantic knowledge). In particular, I have grounded the IS-A relationship in the "plausible entailment" definition of synonymy and hyponymy, carefully distinguished the semantic characterization of selectional constraints from the inferential characterization, and defined the central ideas of the proposal — selectional preference strength and selectional association — in terms of relative entropy, an information-theoretic relationship that is well understood and has a clear, intuitive interpretation.

3.9 Summary

To sum up the main points of this chapter, I reviewed the "semantic" and "inferential" views of selectional constraints, finding empirical problems with the former and practical problems in formalizing the latter. As an alternative, I proposed formalizing selectional constraints in inferential terms, but "hiding" the actual inferential processes within the definition of a conceptual taxonomy by grounding the taxonomy in the notion of "plausible entailment." I then defined selectional constraints in terms of preference, using an information-theoretic relationship between predicates and the taxonomic classes of arguments.

The rest of the chapter was devoted to a discussion of the expected behavior of the model, and a demonstration of actual behavior by means of a computational implementation. Two sets of empirical data were considered: the set of "traditional" examples from the literature that were introduced over the course of the discussion, and stimuli containing "plausible" and "implausible" predicate-argument combinations as determined using ratings tasks with human subjects.

In the chapters that follow, I will consider applications of this model. In Chapter 4, I demonstrate a relationship between selectional preference strength and the argument realization properties of a class of verbs in English, and sketch how the computational model I have proposed might fit into a model of verb acquisition. In Chapter 5, I explore the application of the implemented model to the practical problem of syntactic disambiguation in unconstrained text.

Chapter 4

Selectional Preference and Implicit Objects

In this chapter, I investigate one application of the model proposed in Chapter 3, exploring the relationship between selectional constraints and argument omissibility for verbs in English. It has been observed that the ability of some verbs to omit their objects is connected with the inferability of properties for that argument; I argue inferability can to a great extent be identified with the selectional information carried by the verb. This hypothesis is supported by a computational study: the first experiment demonstrates that verbs permitting implicit objects tend as a group to select more strongly for that argument than obligatorily transitive verbs; the second experiment demonstrates that the tendency in practice to drop the object of verbs correlates with selectional preference strength; and a third experiment investigates the inferability of direct objects for verbs that do and do not require a salient antecedent for that argument in order for it to be omitted. I conclude the chapter with a discussion of some possible implications of this study for accounts of verb acquisition by children.

4.1 Overview

In this chapter, I apply the definition of selectional preference proposed in Chapter 3 to a linguistic problem, namely the question of how it arises that direct objects are optional for some transitive verbs in English and not for others. I begin by defining the syntactic phenomenon of interest, which has sometimes been referred to as intransitivization or object deletion. I restrict my attention to just those omissions that are licensed on the basis of lexical properties of the verb — that is, I am concerned with object omission as a case of diathesis alternation (Levin, 1989). After discussing the relationship between selectional constraints and properties of implicit objects, I develop the hypothesis that strong selectional preference is in fact a requirement for verbs that participate in implicit object alternations, and that strength of selectional preference is connected with how easily properties of arguments can be inferred.

4.2 Implicit Object Alternations

Diathesis alternations are variations in the ways that verbs syntactically realize their arguments. For example, (102) shows an instance of the well-known dative alternation, and (103) shows an instance of the causative-inchoative alternation (Levin, 1989):

(102) a. John gave the book to Mary.
b. John gave Mary the book.
(103) a. John opened the door.
b. The door opened.

Such phenomena are of interest because they stand at the border of syntax and lexical semantics: explaining why a verb expresses its semantic content in a particular syntactic form is part of understanding the nature of its lexical representation.

In this chapter, I focus on a particular set of diathesis alternations having to do with the optionality of direct objects in English. Related terms in the literature include *object deletion*, *intransitivizations*, *null complements*, *implicit objects*, and *optional arguments* (not to be confused with the *implicit arguments* of (Roeper, 1987)). The goal of this section is to define precisely the phenomena with which I am concerned.

Intuitively, the focus is on transitive verbs for which the direct object, when omitted, is nonetheless understood. I will call such omissions *object-drop phenomena*, and the verbs for which they are possible *object-drop verbs*. I will refer to omitted (null, implicit) objects of such verbs as *dropped* or *implicit* objects. In terms of traditional syntactic subcategorization (see, e.g., (Akmajian and Heny, 1975, p. 56ff)), these are the verbs whose subcategorization frames specify an optional NP direct object:

(104)

$$\begin{bmatrix} +\text{V} \\ +[\underline{\quad}\ (\text{NP})] \end{bmatrix}.$$

In descriptions based purely on phrase structure, these are verbs having both a transitive and an intransitive expansion, as is the case for *sing* in the following fragment from (Gazdar et al., 1985, p. 110):

(105) a. VP $\rightarrow$ H[1]
b. die, eat, sing, run, . . .
c. runs
(106) a. VP $\rightarrow$ H[2], NP
b. sing, love, close, prove, . . .
c. prove the theorem

Crucially, it is the lexical representation of such verbs that is taken to license the omission of the direct object. That is, it is important to distinguish between *lexically conditioned* phenomena, which are relevant to this investigation, and *non-lexically conditioned* phenomena, which are not. This distinction is inspired by the distinction that Fellbaum and Kegl (1989) draw between discourse-conditioned and lexically-conditioned intransitivity. Their class of lexically-conditioned intransitivizations is essentially equivalent to Levin's (1989) *indefinite object alternation*, and I have extended it to include what Cote (1992) calls the *specified object alternation*. The latter also appears to be lexically specified, but additionally requires that the context provide a salient antecedent for the null object. (See Section 4.2.3 for details.) Since the

specified object alternation interacts so strongly with discourse context, I have chosen to identify Fellbaum and Kegl's discourse-conditioned cases using the more neutral term *non-lexically conditioned*.

4.2.1 Non-lexically conditioned object omission

Let us consider the difference in more detail, beginning with non-lexically conditioned object omissions. Although English is not noted for its ability to drop the arguments of verbs — as contrasted with Japanese, for example, in which subject-drop and object-drop are both quite frequent — there are in fact many circumstances in which a transitive verb in English may appear without its direct object. Non-lexically conditioned intransitivizations frequently involve habitual, characteristic, or repeated activities or properties of the subject:

(107) a. I thought you said your dog doesn't bite!
b. That is not my dog.

(108) Pussycats eat, but tigers devour.

(109) Religion integrates and unifies. [From the Brown Corpus]

Contrasts or progressions also appear to license the omission of direct objects:

(110) a. You wash, I'll dry.
b. It slices! It dices! [From a TV commercial for the Veg-O-Matic][1]
c. ...the order to load, prepare for action and be on the alert. [From the Collins COBUILD Dictionary]

(111) a. Driver to police officer: If I give you $50, will you ignore this traffic violation?
b. Police officer to driver: You pay, I'll ignore.

Instructions also license this behavior:

(112) a. Lather. Rinse. Repeat.
b. Bake for an hour at 350°.

Phenomena of this kind do not appear overly sensitive to the particular verb: a context in which the direct object is omissible can be constructed for just about any transitive verb, by creating a situation in which the verb is interpreted within one of the above licensing contexts. This is what leads to the conclusion that non-lexically conditioned phenomena are an issue of grammar (and discourse), and not a matter of lexical representation.

4.2.2 Lexically-conditioned object omission

Even when lexically unconstrained syntactic or discourse processes are excluded from consideration, there are still numerous ways in which transitive verbs in English can specify the optionality of their complements. These can be distinguished along three dimensions of the omitted argument: syntactic category, definiteness, and semantic type.[2]

[1] Informants comment that "It slices! It dices! It devours!" would be equally good.

[2] Unlike (Grimshaw, 1979; Pesetsky, 1982), I will not be considering at all the more general case of *predicates* appearing without their complements, as in

The first dimension is syntactic: lexically-conditioned object-drop phenomena may involve the omission of a sentential argument (113), an NP argument (114), or a PP argument (115):

(113) a. I knew that the money had been stolen.
b. I knew.
(114) a. I called my mother.
b. I called.
(115) a. I contributed ten dollars to the emergency fund.
b. I contributed ten dollars.

Fillmore (1986) distinguishes finer grammatical types such as indicative that-clause direct objects and subjunctive that-clause direct objects — e.g. complements of *know that...* versus *insist that...* — but the coarse three-way distinction according to constituent type suffices for present purposes.

Second, the dropped object may have an *indefinite* interpretation (117a) or a *definite* interpretation (117b). In Fillmore's terms, this is the distinction between indefinite null complements and definite null complements.

(116) What did John do at noon?
(117) a. He ate (though I'm not sure what he ate).
b. He called (#though I'm not sure who he called).

Fillmore comments that a useful test for the distinction is "whether it would sound odd for a speaker to admit ignorance of the identity of the referent of the missing phrase," as is presumably the case for the questionable continuation in (117b).

Third, the semantic type of the object can involve truth conditions, or it can involve an entity or entities. (Roughly speaking, these correspond to the types t and e, respectively, found in model-theoretic semantics — see, e.g., (Dowty, Wall, and Peters, 1981).) Arguments involving truth conditions may be interrogative (118), exclamatory (119), or propositional (120) (Grimshaw, 1979):

(118) a. I wonder how fast Bill can run.
b. I wonder.
(119) a. I know how very fast Bill can run.
b. I know.
(120) a. I forgot that Bill is a runner.
b. I forgot.

Arguments involving entities are usually expressed syntactically using noun phrases, as one would expect:

(121) a. Bill read a magazine.
b. Bill read.

Although the syntax and semantics of the dropped object seem closely related, Grimshaw argues convincingly that they must be distinguished when specifying the selectional properties of a verb. She supports her claim using the phenomenon of control by concealed questions. The decisive example is reproduced here as (122) (from Grimshaw's (113, 114)):

(a) It's amazing how quickly Bill can run.

(b) It's amazing.

Nor will I be considering a possible fourth dimension of variation, namely whether or not the null argument is projected at the level of syntactic structure (see (Rizzi, 1986)).

(122) a. Bill asked me the time, so I inquired.
b. Bill asked me the time, so I inquired what the time was.
c. *Bill asked me the time, so I inquired the time.

The concealed question in the first clause of (122a) controls the interpretation of the implicit argument in the second clause. The contrast between (122b) and (122c) shows clearly that the control relationship holds at the level of semantic type and not syntactic form: the direct objects of *ask* and *inquire* in (122c) are identical with respect to syntactic form, and therefore, since the sentence is ungrammatical, it cannot be the case that *inquire* selects its direct object on purely syntactic grounds.

Within this three-dimensional space, only a subset of the phenomena will be considered here. In terms of the dimensions just defined, the behavior I will describe as "optionality of direct objects" can be characterized as follows:

- syntactic type NP
- either definite or indefinite
- semantic type e (entity or entities).

Though only a subset of the full range of null complements, this description still leaves a fair amount of ground to cover; in addition, any progress made toward accounting for this subset of phenomena can serve as a starting point in efforts to account for the rest.[3]

From this point on, then, the term *object-drop phenomena* will refer only to those (lexically-conditioned) phenomena that fit within the dimensions just given, and *object-drop verb* and *non-object-drop verb* should be understood accordingly. The remainder of this section is concerned with diagnostics that determine whether or not a particular verb should be considered an object-drop verb.

4.2.3 Diagnostics

The class of object-drop phenomena corresponds to the union of two diathesis alternations: the indefinite object alternation (IOA) of (Levin, 1989), and the specified object alternation (SOA) proposed by (Cote, 1992). For present purposes, therefore, the linguistic diagnostics used by those authors to characterize the alternations can be used to demarcate the boundary between object-drop verbs and non-object-drop verbs.

Cote suggests three diagnostics to determine when a verb participates in the indefinite object alternation. First, the verb cannot have a null object when a salient antecedent is present.

(123) a. Did Cheetah eat all the bananas?
b. #Yes, he ate.

Second, the verb's appearance with a null object can cancel apparent equivalence with an antecedent.

(124) a. Did Cheetah eat the bananas?
b. He ate, but not the bananas. He had mangos instead.

Third, the verb's appearance with a null object can introduce a new entity into the discourse context.

[3] Pustejovsky (1991, footnote 13) briefly considers the case of dative PP omission, conjecturing that the omission of PP arguments (as in *Cordelia told the story*) is related to the semantic "connectedness" of the verb-object combination, accounting for the contrast between *Cordelia told the story* and **Cordelia told the secret.* It may be possible to develop Pustejovsky's conjecture further using the argument presented here.

(125) a. Did you cook today?
b. Yes, and it came out delicious.

As noted above, Fillmore (1986) suggests a diagnostic (also adopted by (Rispoli, 1992)) to distinguish the indefinite object cases from the definite (specified) object cases: for the former, the speaker must be understood to have a specific antecedent in mind, and it is infelicitous to indicate otherwise.

(126) a. When I peeked into John's room he was reading;
b. now I wonder what he was reading.
(127) a. When I peeked into John's room he was winning;
b. #now I wonder what he was winning.

A corresponding diagnostic for picking out verbs that *can* take specific null objects is to construct a discourse context where the antecedent is clearly salient.

(128) a. Remember that game we were discussing?
b. Well, John won, and he bragged about it all night.
(129) a. Remember that door we were having trouble with?
b. *Well, John unlocked, and he promised to make me a copy of the key.

Such a context must be constructed with care, however, in order to avoid creating a discourse context that supports *non*-lexically conditioned object-drop phenomena such as those discussed in Section 4.2.1. For example, (130) should not be considered evidence that *lift* participates in the specified object alternation:

(130) a. John and Bill will go from bedroom to bedroom looking under mattresses for hidden money.
b. John will lift and Bill will look.

Because the judgements are sometimes subtle, it is helpful to use a dictionary as a point of reference. I have used the Collins COBUILD English Language Dictionary (Sinclair (ed.), 1987), which has the convenient property of organizing verb subcategorization information according to verb sense. Thus it is possible to identify a verb as a likely participant in implicit object alternations simply by seeing whether some (non-marginal) sense of the verb is annotated with both V and V+O; after which, of course, one can apply further diagnostics. If no sense of a verb permits both the V and V+O frames in this dictionary, it can reliably be excluded from consideration.

4.2.4 Properties of implicit objects

Although the participation of a verb in implicit object alternations is usually encoded simply as a set of structural alternatives — as illustrated by (104) and (105) — it is clear that the alternation has implications for interpretation, as well. In particular, when a verb's object is dropped, the missing argument is taken to have the properties of "prototypical" objects of the verb. For example, Levin (1989, p. 7) gives the following example in characterizing the indefinite object alternation (her examples (17a,b)):

(131) a. Mike ate the cake.
b. Mike ate. (→ Mike ate something one typically eats),

and Ellen Prince points out that for many people, (132) is natural under a "washing dishes" interpretation,

(132) You wash and I'll dry!

but humorous in a situation where the speaker and listener are giving the baby a bath.[4]

These property inferences are, in fact, very much like the inferences drawn when the argument position is occupied but unspecific. Fodor (1977) comments that selection restrictions have been used not only to predict semantic anomaly, but also to provide inferences of this kind, filling out the meaning of sentences containing a pronoun in argument position.

(133) a. Sincerity admires John.
b. This one admires John.

She writes:

> In [(133b)], the subject noun phrase is not specified for animateness, so there is no direct conflict with the selection restriction on the verb *admire* which requires its subject to be animate (or more precisely, to be capable of higher psychological functions). But the selection restriction on the verb induces an interpretation of the subject as if it WERE an animate noun phrase. (p. 195)

This relationship between selectional constraints and omitted (or underspecified) arguments appears in the discussions of a number of authors about lexical representations and how they might be used in processing. Jackendoff's (1990, p. 52) discussion of optional complements is one such instance. Within his representational scheme, the lexical entry for a verb specifies (i) the syntactic form of its complement — for example, a subcategorization frame — together with (ii) some expression of semantic selection for that complement, (iii) a lexical conceptual structure having open argument positions, and (iv) an annotated distinction between obligatory arguments and those that are optional. A selectional restriction is considered to be a part of the verb's lexical conceptual structure. For example, the lexical entry for *drink* includes a selectional restriction on the direct object (more precisely, on the argument position within the lexical conceptual structure) in the form of the conceptual annotation LIQUID.

Jackendoff suggests that in processing, selectional restrictions are enforced not by means of an independent filter, but rather via a mechanism he calls *argument fusion*: when the lexical conceptual structure associated with the verb is combined with an overt argument, the conceptual content of the argument is combined with the (partial) conceptual content already found in the argument position. Should there be a clash of types — for example, a non-liquid direct object for *drink* — this fusion cannot take place, and a selectional violation results. Should an optional argument *not* be overtly specified, the empty argument of the verb will nonetheless be attributed with appropriate "default" features by virtue of the partial conceptual information within the verb's lexical entry. Rizzi (1986, footnote 6), in a similar vein, suggests that optional arguments correspond to thematic roles that are "saturated" in the lexicon rather than syntactically. Common to these discussions is the idea that selectional information specified in a verb's lexical entry is combined with objects when they are overt, and is ascribed to those arguments when they are omitted.

The evident relationship between selectional constraints and property inferences suggests the following hypothesis: verbs that permit implicit objects select strongly for that argument. This makes sense on intuitive grounds — relevant properties of omitted arguments are clearly inferred somehow, and the verb seems the most likely place to look for the relevant information. To state the hypothesis another way, if a verb does *not* carry sufficient selectional information to permit the relevant object properties to be inferred, then it should not permit that argument to be omitted.

The following examples make the intuition behind this hypothesis quite clear.

[4] Personal communication; she attributes the example to Gregory Ward.

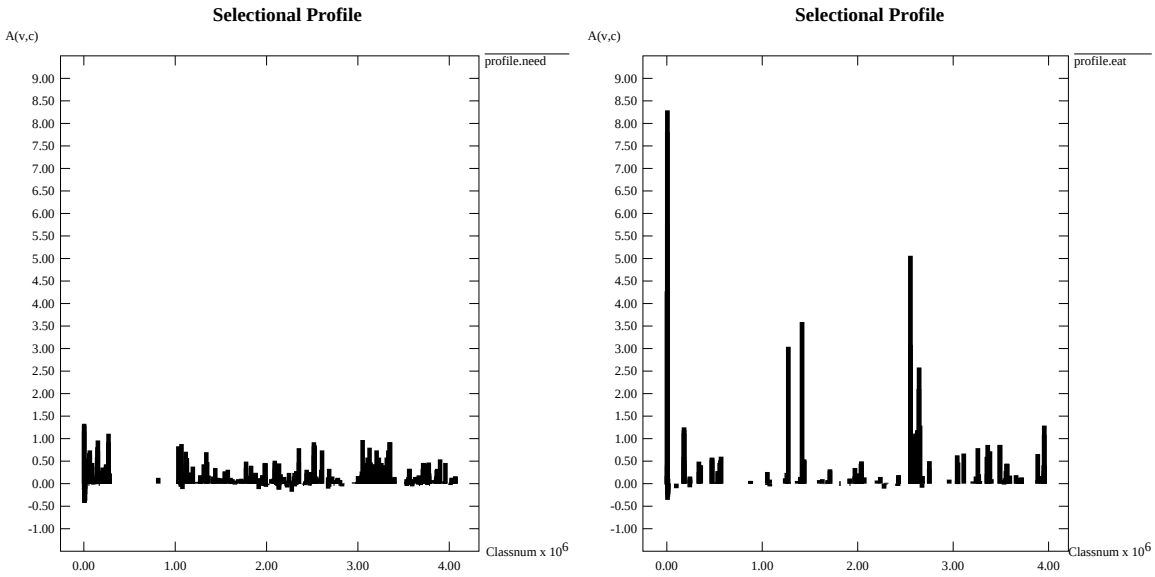


Figure 4.1: Selectional behavior of *need* and *eat*

(134) a. John ate something.
b. John ate food.
c. John ate cereal.

(135) a. John needed something.
b. John needed assistance.
c. John needed help with his homework.

In (134b), the direct object seems to contribute very little additional information over and above (134a). In example (135b), the direct object seems to contribute more information; after all, John might need any number of things (a ride to work, a haircut, a new computer password), only some of which are kinds of assistance. One can imagine how a language might incorporate such informational differences into its syntactic behavior. In cases like (134b), but not (135b), the direct object contributes so little information that often it might as well be omitted entirely, and this behavior ultimately becomes incorporated into the subcategorization of the verb.

The connection between inferability of the direct object and its omissibility is not a new one — similar observations have been made by (Lehrer, 1970; Rice, 1988; Fellbaum and Kegl, 1989). However, the formalization of selectional preference proposed in Chapter 3 provides a necessary link between inferred information and selectional properties of the verb, and provides a new and formal interpretation of what information is. Furthermore, unlike traditional characterizations of selectional constraints as sortal restrictions, the information-theoretic proposal makes it possible to discuss selection quantitatively rather than in all-or-nothing terms. Rather than suggesting that *need* provides no selectional information at all about its direct objects, one need only claim that the information provided is comparatively less than for some other verbs.

This is illustrated in Figure 4.1, which shows the selectional behavior of the verbs *need* and *eat* as determined experimentally from the Brown corpus, using the computational model proposed in Chapter 3.

The taxonomic classes from WordNet have been laid out sequentially along the horizontal axis, with each class taking the x-value corresponding to its unique numerical identifier in WordNet.[5] The vertical axis indicates the magnitude of selectional association between the verb and each concept.[6] As the figure shows, the verb *need* (profile at left) does constrain its direct object to some extent, having a stronger selectional association with some concepts than others, although the overall effect is relatively small in magnitude and not conceptually specific. However, the selectional behavior of *eat* (profile at right) is markedly different: qualitatively its pattern of selectional association is far more specific (the highest peak corresponds to the WordNet class ⟨food⟩), and quantitatively its overall selectional preference strength is much greater. The ability to quantify selectional constraints in this way makes it possible to put a precise formulation of the hypothesis under discussion to an empirical test.

4.3 Experiment 1: Selection and Optionality

4.3.1 Procedure

Several computational experiments were carried out in order to test the hypothesis that object-drop verbs can be distinguished from non-object-drop verbs on the basis of selectional preference strength. The general procedure was as follows:

- A sample of verbs was chosen, comprising a set of 34 verbs that occur frequently in parental speech to children.[7]
- Each verb was classified as object-drop or non-object-drop. A verb was classified as object-drop only if (a) some sense of the verb is annotated with both V and V+O in (Sinclair (ed.), 1987), and (b) that sense is "close enough" to the central meaning of the verb, as opposed to an extremely specialized sense. The latter criterion is a question of personal judgement: some sense of each verb in (136) and (137) permits both subcategorizations, but in cases like (137) I decided the senses permitting the alternation were too specialized to warrant categorizing the verbs as object-drop.

(136) a. John called (someone) at 3pm.
b. John packed (a suitcase) quickly before leaving.
c. John stole (some money) and was caught.

(137) a. John opened (a discussion) with a question.
b. John showed (a work of art) in New York.
c. The missile hit (a target) and exploded.

- For each verb, the selectional preference strength was calculated as described in Chapter 3. The experiment was replicated using several different sources for verb-object co-occurrence frequencies; see details below.[8]
- Statistical tests were carried out to see if the object-drop verbs did in fact have higher strengths of selectional preference than their non-object-drop counterparts, as predicted by the hypothesis.

[5] These unique identifiers correspond to positions within a WordNet 1.2 data file; its size is about 4 megabytes.
[6] Values of selectional association have been multiplied by 100.
[7] I am grateful to Annie Lederer for providing this list.
[8] Recall from Section 3.7.1 that verb senses were not distinguished, although sense distinctions may be relevant (see discussion in (Fillmore, 1986)).

In using this procedure to test the hypothesis, the WordNet noun taxonomy is assumed to be a reasonable representation of the conceptual taxonomy available to a language user. The psycholinguistic considerations underlying the taxonomy are presented by Miller (1990a); among others, he includes the following:

1. Clinical observations of patients with anomic aphasia lend support to the isolation of nouns into a separate lexical subsystem.

2. A hierarchical organization of the noun lexicon is supported by psycholinguistic evidence concerning anaphoric nouns and comparative constructions. (E.g., *He owned a rifle, but the gun had not been fired*; *#A rifle is safer than a gun/#A gun is safer than a rifle*.)

3. The noun hierarchy strongly reflects function. (Miller comments, "At least since Dunker (1945) described functional fixedness, psychologists have been aware that uses to which a thing is normally put are a central part of a person's conception of that thing.")

In (Miller et al., 1990), Miller makes the following general comment:

> Beginning with word association studies at the turn of the century and continuing down to the sophisticated experimental tasks of the last twenty years, psycholinguists have discovered many synchronic properties of the mental lexicon that can be exploited in lexicography . . . Inasmuch as it instantiates hypotheses based on results of psycholinguistic research, WordNet can be said to be a dictionary based on psycholinguistic principles.

As a second assumption, the verb-object samples used in the experiment are taken to be representative of actual usage. In order to ensure that this is the case, the corpora used were as balanced as possible. One experiment used the Brown Corpus (Francis and Kučera, 1982), which, though smaller than some of the text corpora now available (about a million words, total), is the largest readily available sample of English text explicitly designed to be balanced across genres. A second experiment used parental speech from the CHILDES corpus (MacWhinney, 1991), which is to my knowledge the largest and broadest sample of parent/child interaction available. A third experiment used data collected by Annie Lederer in an unpublished study of verb-object norms.

It is very important to note that no statistics at all were collected concerning the frequency with which verbs in the study do or do not appear with omitted arguments. Only *overt* verb-object co-occurrences go into the estimation of selectional preference strength; therefore the independent measure is not tainted by information about the property it is being used to predict.

4.3.2 Results

Brown Corpus. The first version of the experiment used a sample of 33,136 verb-object co-occurrences extracted from the parsed version of the Brown Corpus in the Penn Treebank (Marcus, Santorini, and Marcinkiewicz, 1993). The Treebank parses encode only surface objects, and since pronoun-antecedent relationships are not encoded, pronouns in object position were ignored. In multiple-noun compounds, the last noun (reliably the head of the NP) was taken to be the direct object.[9]

[9] I am very grateful to Rich Pito for his TGREP utility, which made it possible to search for and and extract specific structural patterns from the Penn Treebank.

In this version of the experiment the object-drop verbs have significantly stronger selectional preference strength than the non-object-drop verbs according to a Mann-Whitney U test (Milton, 1964) (N1=15, mean1=2.97, stdev1=0.98, N2=19, mean2=1.73, stdev2=0.93, U=55, $p = .001$).

CHILDES. The second version of the experiment used a sample of 34,710 verb-object co-occurrences extracted from parental speech in the CHILDES (Child Language Data Exchange System) collection of parent-child interactions (MacWhinney and Snow, 1985; Sokolov and Snow, to appear). The CHILDES data are not parsed, so in order to identify the direct objects of verbs, the parental turns from the CHILDES data were extracted and run through a probabilistic part-of-speech tagger, and direct objects identified using a simple heuristic procedure. Essentially the procedure looked to the right of the verb for either a noun, a sequence of nouns (in which case the last one was identified as the head), or another part of speech such as a preposition or adverb that would suggest terminating the search without identifying an object. Although this procedure led to a noisy sample, I inspected a sub-sample by hand and judged the results to be reasonable; in addition, an earlier version of the experiment applied the same heuristic to the Brown Corpus (before it was available in parsed form) and the results were essentially the same as those just described.[10]

In this second version of the experiment, the object-drop verbs again have significantly stronger selectional preference strength than the non-object-drop verbs, according to a Mann-Whitney U test (N1=15, mean1=2.25, stdev1=0.94, N2=17, mean2=1.13, stdev2=0.64, U=37, $p < .0005$). (Verbs *do* and *have* were excluded from the heuristic object-finding procedure owing to their use as both verbs and auxiliaries, hence the experiment included 32 rather than 34 verbs.)

Human subject data. The third version of the experiment used a sample of 2,655 verb-object co-occurrences collected in an unpublished norming experiment by Annie Lederer. Ten subjects were instructed to name "the top ten things that you..." for each of the 34 verbs — something very much along the lines of the "Family Feud" television game show. They were told to restrict themselves to one-word answers, and to list fewer items if ten did not come to mind easily. In some cases, subjects gave two-word responses despite the instructions (e.g. *pour orange juice*, *open car door*); in adapting the norms to the experiment reported here those responses were excluded.

In this third version of the experiment, the object-drop verbs once again have significantly stronger selectional preference strength than the non-object-drop verbs, according to a Mann-Whitney U test (N1=15, mean1=2.17, stdev1=0.42, N2=19, mean2=1.66, stdev=0.42, U=57, $p < .0025$).

4.3.3 Discussion

The results of Experiment 1 confirm the hypothesis that verbs participating in the implicit object alternation select more strongly for their direct objects than verbs that do not. Replications using several different corpora to estimate verb-object co-occurrences lend the result additional credibility: the difference is apparently not the result of some quirky statistical behavior in a particular corpus.

It is important to note, however, that there is no clear threshold separating the two groups of verbs. For example, using the Brown Corpus data, the three "weakest" object-drop verbs are *call*, *hear*, and *watch*,

[10] All parental data available in CHILDES were merged; these included data gathered by the following researchers: Bates, Bernstein, Bloom, Bohannon, Braine, Brown, Clark, Evans, Garvey, Gathercole, Gleason, Hall, Higginson, Howe, Kuczaj, MacWhinney, Sachs, Snow, Suppes, Vanhouten, and Warren. See (MacWhinney and Snow, 1985) for details. I am grateful to Eric Brill for his assistance in tagging the CHILDES data.

with selectional preference strengths ranging from 1.52 to 1.97; the three "strongest" non-object drop verbs are *hang*, *wear*, and *open*, with selectional preference strengths ranging from 2.93 to 3.35. The results using other corpora show similar behavior. This might mean that selectional preference strength is being poorly estimated in some cases, or it might mean that there are other factors involved in determining whether the direct object is optional.

There is some evidence to suggest that usage biases may be leading to inaccurate models of verb selection in some cases. For example, in the Penn Treebank parses for the Brown Corpus the direct object distribution for the verb *say* is systematically contaminated by time adverbials, as in the following:

(138) a. He still [VP says [NP every day] again: "Let there be light"]!
b. Governor Notte [VP said [NP last night]...]

This is one likely explanation for the selectional preference strength of 2.82 for *say* as estimated from this corpus. In addition, inappropriate word senses appear to be having undue influence for some low-frequency verbs. For example, *governor*, *head*, *official* and *tool* together account for 6 of the 11 direct object instances observed for the verb *hang*; these can all be grouped together under the heading ⟨person,3174⟩ if *head* is interpreted in its sense as CHIEF and *tool* is interpreted as PUPPET or SLAVE. If word senses were disambiguated, the co-occurrence *hang head* would contribute probability to senses such as BODY_PART and *tool* would be associated with IMPLEMENT and the like, and the overall selectional preference strength would be lower.

Despite these biases, I am inclined to take the latter position, namely that, even if the estimated values for selectional preference strength were completely accurate, selectional preference would not completely account for omissibility of objects. I take up other factors that might be involved in the general discussion.

4.4 Experiment 2: Selection and Frequency of Omission

The previous experiment investigated the hypothesis that optionality of the direct object is connected to selectional preference, the rationale being that strength of selectional preference is, as formalized here, a measure of how easy it is to infer or reconstruct necessary properties of the omitted object. Although the results do not support a categorical distinction solely on the basis of selectional properties, they do show that selectional properties are relevant to lexically-specific syntactic behavior.

Given that selection is relevant to lexical-syntactic properties — that is, lexical knowledge bearing on syntactic competence — a natural question to ask is whether selectional preference affects syntactic performance, as well. In particular, if selectional preference strength measures how much information a verb carries about its object, then properties of omitted objects should in some sense be more easily inferred for strongly rather than weakly selecting verbs.

Ease of inference is a subject for investigation by psycholinguistic rather than computational methods. However, in performance, a speaker or writer is likely to be influenced by how easy it will be for the listener or reader to arrive at the correct interpretation. In particular, one would expect that verbs for which the object is readily inferable will omit that argument correspondingly more frequently than verbs for which the object is not easily inferred. In a second experiment, therefore, I have again associated ease of inference with strength of selectional preference, this time exploring the connection between selectional preference and the omission of direct objects in actual performance.

4.4.1 Procedure

In order to determine the frequency with which verbs omit their objects, I extracted from the Brown Corpus a sample of 100 instances of each verb used in the preceding experiment (or as many instances as were available, if fewer). For each instance, I used the full sentence in which the verb appeared, together with the full preceding sentence, to decide whether or not this instance was an example of an implicit object construction. The judgements were made on the basis of intuition, together with the linguistic diagnostics discussed in Section 4.2.3.

(139) a. In the fullness of her vocal splendor, however, she could sing the famous scene magnificently.
b. Altogether fifteen virtually unknown Rodgers and Hart songs are sung by a quintet of able vocalists.
c. Rouben Ter-Arutunian, in his stage settings, often uses the scrim curtain behind which Mr. Cole has placed couples or groups who sing and set the mood for the scenes which are to follow.

For example, (139c) was counted as containing an implicit argument instance for *sing*, and (139a) and (139b) were not.

In this process, it was more convenient to use the original part-of-speech tagged Brown Corpus rather than the parsed version found in the Penn Treebank, though the latter was still used for estimating selectional preference strength. Since in the original Brown Corpus the uses of *have* as auxiliary and verb are not distinguished, that verb was excluded from the sample, leaving the other 33 verbs from Experiment 1. Selectional preference strength was determined for each verb in exactly the same fashion described earlier, for each of the same three corpora.

4.4.2 Results

A correlation between selectional preference strength and object omissions emerged in each of the three versions of the experiment: Brown Corpus (N=33, r=.48, F(1,31)=9.53, $p < .01$, $p(F) < .005$), CHILDES (N=32, r=.36, $p < .05$, F(1,30)=4.33, $p(F) < .05$), human subject data (N=33, r=.58, $p < .001$, F(1,31)=15.74, $p(F) < .0005$). Figure 4.2 shows a plot of the relationship using the human subject data — *Strength* refers to selectional preference strength, and *Implicit* is the proportion (between 0.0 and 1.0) of instances appearing with an implicit object.

Although some verbs deviate by failing to omit their objects despite very strong selection for the direct object, it is interesting to notice that the converse does not hold: verbs do not omit their objects frequently unless they possess a high selectional preference strength. I would argue that this pattern reflects an underlying hard requirement, namely that strong selection is a necessary condition for object omission. Whatever other sources of information may be available for inferring properties of implicit objects, selectional information carried by the verb is a prerequisite.

4.5 Experiment 3: Distinguishing Subclasses of Object-drop Verbs

In Experiment 1, verbs participating in the indefinite object alternation (IOA) and the specified object alternation (SOA) were combined into a single group. One might predict, however, that the selectional properties of verbs in the two subclasses might differ — if a verb requires that an antecedent be available

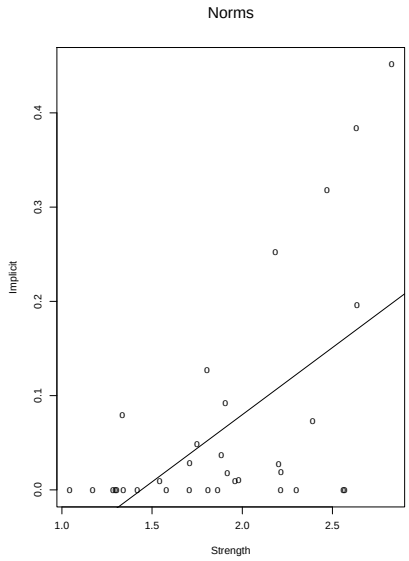


Figure 4.2: Correlation between selection and implicit objects

IOA	SOA
drink	call
eat	explain
pack	hear
read	play
sing	pour
steal	pull
write	push
	watch

Table 4.1: Subclassification of object-drop verbs

in the discourse context, the verb itself might not contribute as much information about the omitted object. This prediction can be tested using the same procedure as in Experiment 1, replacing the object-drop and non-object-drop groups with groups of verbs participating in the indefinite and specified object alternations, respectively.

Table 4.1 shows the division of the object-drop verbs from the sample into those two subclasses. Testing the predicted difference in selectional preference strength between the two classes yielded the following results, organized, as before, according to the corpus according to which selectional preference strength was estimated. (Group 1: IOA; group 2: SOA.)

Brown Corpus: N1=7, mean1=3.43, stdev1=0.75, N2=8, mean2=2.57, stdev2=1.02, U=13, $p = .05$.

CHILDES: N1=7, mean1=2.51, stdev1=0.80, N2=8, mean2=2.03, stdev2=1.05, U=14, $p = .1$.

Human subject data: N1=7, mean1=2.14, stdev1=0.54, N2=8, mean2=2.20, stdev2=0.33, U=26, n.s.

Although these results might optimistically be interpreted as supporting the hypothesis, the difference between the two groups has not been convincingly demonstrated. One likely explanation is that the the two groups — respectively containing 7 and 8 verbs — are too small.

In the interest of obtaining results from a larger sample of verbs, I repeated the experiment using the verb classification in (Lehrer, 1970). Lehrer divides verbs permitting "object deletion" into four types:

I. Verbs that imply highly specific semantic objects, with the identity of the object not affected by discourse context (e.g. *eat*, *read*).

II. Verbs that imply two or more fairly specific objects; for these the discourse might have an effect on how the missing object is interpreted.

III. Verbs that allow the object to be deleted in certain discourse contexts (i.e. when an antecedent is easily identified) without loss of meaning. This group is mutually exclusive with type I, and overlaps with type II.

IV. All remaining verbs that permit object deletion (e.g. *steal*, *interrupt*). These tend not to be associated with a highly specific object, and "sometimes behave as type III verbs do, but less regularly."

There seems to be a fairly clear correspondence between Lehrer's Type I and the implicit object alternation and between Type III and the specified object alternation (see Appendix B.7). Lehrer writes:

> [Type III verbs] allow their objects to be deleted when the object has appeared in the preceding discourse. Or, looking at the matter from an analytic point of view, the following verbs tend to 'pick up' objects from the preceding discourse.

Her types II and IV were excluded, since their status with regard to the indefinite/specified distinction is unclear. Using the Brown Corpus to estimate selectional preference strength, Lehrer's Type I verbs select more strongly for their objects than the Type III verbs, according to a Mann Whitney U test (N1=34, mean1=4.46, stdev1=1.75, N2=42, mean2=3.04, stdev2=1.29, U=356, $p < .0001$). However, although there is a difference when the experiment is done using the CHILDES corpus, it is not significant (N1=27, mean1=3.32, stdev=1.73 N2=31, mean2=2.86, stdev=1.49, U=359, n.s.). The empirical evidence is therefore suggestive, but not conclusive. The difference between the behavior of the Brown Corpus and CHILDES samples may result from the fact that they respectively contain edited text and spontaneous speech, but I have not yet investigated this possibility in detail.

4.6 General Discussion

The experimental results confirm that there is an interaction between strength of selectional preference and the optionality of direct objects. Experiment 1 shows that verbs permitting the omission of their direct objects select for that argument more strongly than do verbs for which the object is obligatory. Experiment 2 lends support to the claim that this relationship is a causal one, showing that in actual performance there is a correlation between selectional information and the frequency with which the reader or listener is expected to infer an omitted object in practice. Finally, Experiment 3 investigated the claim that selectional information is connected to ease of inference. It suggests that verbs with external sources of information — salient antecedents in the discourse — may select less strongly than verbs for which that information is unavailable.

As I pointed out earlier, however, selectional preference information alone is not enough to provide a categorical distinction between verbs that do and do not drop their objects. In the remainder of this section I discuss several other factors that may be connected with the indefinite object alternation, among them aspectual constraints and taxonomic relationships in the lexicon.

4.6.1 Aspectual constraints

The account of implicit object alternations presented here stands in contrast to the more common view of diathesis alternations as being closely tied to some specific aspect of the verb's semantic content. For example, the dative alternation is related to the element of transfer.

(140) a. John gave the book to Mary.
b. John gave Mary the book.

(141) a. John told the story to Mary.
b. John told Mary the story.

As Pinker (1989) puts it,

> Dativizable verbs have a semantic property in common: they must be capable of denoting prospective possession of the referent of the second object by the referent of the first object . . . verbs of communication are treated as denoting the transfer of messages or stimuli, which the recipient metaphorically possesses. (p. 48)

Pinker goes on to discuss semantic constraints on a range of other argument structure alternations, including causatives, locatives (spray/load), and passives.

Although verbs permitting implicit objects do not appear to be linked by any factor so tightly bound to their semantics, there are some factors connected to aspect that are relevant. These are most easily described in terms of Vendler's (1967) aspectual classes, and their related time schemata. To describe these aspectual distinctions very briefly, *activities* and *accomplishments* admit continuous tenses, whereas *achievements* and *states* do not — the former but not the latter are processes, in the sense of "successive phases following one another in time" (p. 99). Activities and accomplishments are distinguished by the notion of an end or climax; achievements and states are distinguished by the determinacy of the time period involved. A fuller discussion can be found in (Vendler, 1967); for present purposes, the following diagnostics for distinguishing the four classes will suffice:

- Activity from accomplishment:

(142) a. For how long did he push the cart? [activity]
b. How long did it take him to write the letter? [accomplishment]

- Accomplishment from state:

(143) a. I am writing the letter. [accomplishment]
b. *I am knowing the answer. [state]

- State from achievement:

(144) a. How long did it take to recognize the painting? [achievement]

b. *How long did it take to know the answer? [state]

- Achievement from activity:

(145) a. For how long did he push the cart? [activity]
b. At what time did he recognize the painting? [achievement]

- Activity from state:

(146) a. I am pushing the cart. [activity]
b. *I am knowing the answer. [state]

- Achievement from accomplishment:[11]

(147) a. I wrote the letter without interruption. [accomplishment]
b. *I recognized the painting without interruption. [achievement]

A generalization that appears to hold for verbs participating in the indefinite object alternation is that they describe processes, specifically accomplishments, when used transitively. Most of Vendler's diagnostics for accomplishment verbs are captured concisely in (148):

(148) a. What was John doing?
b. John was drinking his coffee.
c. It took John ten minutes to drink his coffee,
d. and he drank it without interruption.

Furthermore, Mittwoch (1971; 1982) argues that verbs with omitted indefinite objects are interpreted as activities. In (Mittwoch, 1971) she shows that when *drink* appears without a specified object, or with an object of indefinite quantity, as in *drank beer*, the VP must be interpreted as describing an event that has not necessarily been completed.

(149) a. John drank (beer).
b. *John drank up (beer).
c. John drank up the glass of beer.
(150) a. John drank (beer) for two hours.
b. *John drank (beer) in two hours.

This observation accounts for the exclusion from the alternation of verb phrases that have only a completed-event reading. Since the particle *up* in phrases like *drink up* and *eat up* contributes the semantic feature [+completive], it is incompatible with an unspecified interpretation of the "deleted" NP.

Mittwoch's argument suggests a constraint excluding not only verbs appearing with completive particles, but also verbs that carry the [+completive] feature themselves. Browne (1971) makes a similar point, phrased in terms of goal-directness. He points out that a common feature of the verbs in (151), which are prohibited from omitting their objects,

(151) a. *Bill devised.

[11] Vendler does not himself propose this diagnostic, but *without interruption* contexts do seem to capture what he has in mind. The crucial point is even though an achievement can be described as taking a certain amount of time ("It took three hours to reach the summit"), it does not imply that the described action (reaching the summit) took place at each moment during that period.

b. *John consumed/devoured.[12]
c. *Fred debitted.
d. *Max halved (i.e. cut something in two)
e. *The FBI detected.
f. *Selma ignited (i.e. set something on fire)
g. *Moishe exploited.

is that they all presuppose "progress toward an end point at which an idea is existent, the drink is completely gone, the object lies cleft, the FBI is in possession of information, etc."

To summarize, it appears that verbs with indefinite implicit objects are accomplishments when an object is included and activities when it is omitted; furthermore, the indefinite object alternation must not result in a verb phrase that necessarily describes an achievement (completed action, end state).[13] However, not every verb meeting these criteria permits indefinite objects to be omitted. For example, *record* (in the sense of recording sounds on tape) fits the aspectual criteria, but fails to omit indefinite objects.

With verbs permitting the omission of specified rather than indefinite objects, the connections to aspect become less clear.

(152) a. Bill had absolutely no idea what the answer was.
b. John knew, because he'd just looked it up.
(153) a. The ACL conference was in Ohio this year.
b. About four hundred people attended.

For example, *know* is licensed in (152b) despite the fact that it denotes a state, is non-completive, and is not in any sense goal-directed. In (153b), *attend* is, according to the diagnostics, an activity. However, Fillmore (1986) comments that

> [It] is particularly striking that the semantic role of Patient (or Theme) appears not to occur among the definite omissibles. That is, we found no cases of [definite null complements] with change-of-state verbs like BREAK, BEND, CREATE, DESTROY, MOVE, LIFT, and the like. (p. 104)

and it is interesting to note that, in addition to the common thematic role involved, all these verbs are inherently completive.

As a final thought along these lines, aspect is only one of a number of the properties associated with high transitivity that may be associated with failure to omit objects. Hopper and Thompson (1980) identify degree of transitivity with set of parameters that includes aspect (telic or atelic), punctuality (whether or not an action is inherently on-going), volitionality of the agent, affirmation (positive or negative), mode (correspondence to an actual event), agency, and affectedness and individuation of the object. Their analysis may be a useful starting point for an account of implicit object alternations that takes into account not only lexical properties but also features of the discourse context.

[12] As it happens, the points about aspect made here would account for the classic contrast between *eat* and *devour*: the latter, but not the former, entails a completed event of consuming the object. The first definition for *devour* in the American Heritage Dictionary, "to eat up greedily," brings in not just manner but also the completive particle. However, it worth noting that the claim made in this chapter also applies. In an informal experiment, subjects who were asked to produce sentences containing the verb *devour* frequently responded with non-food objects like *book*, *opponent*, and *savings*. If those informally gathered verb-object co-occurrence counts are added to the verb-object norms, then *devour* turns out to have a much weaker selectional preference strength for its object than *eat*.

[13] This last constraint may suggest that *steal*, which I categorized with the IOA verbs, should be excluded from the IOA verbs. Although (Sinclair (ed.), 1987) lists *steal* as permitting frames V or V+O in its core sense, the example they give (*Children often steal*) clearly falls into the category of non-lexically-conditioned object omission.

4.6.2 Taxonomic relationships

Fellbaum and Kegl (1989) adopt the aspectual analysis made by Mittwoch, describing her distinction as one between telic and atelic interpretations. They go beyond aspect, however, to propose an account of the indefinite object alternation phrased in terms of a taxonomic organization of verbs that is analogous to an IS-A taxonomy of nouns.

Central to Fellbaum and Kegl's account is a distinction between two different kinds of IS-A relationships between verbs. They point out that although "nibbling is a kind of eating" and "dining is a kind of eating," the relationship in the former but not the latter case involves manner. An analysis of the set of verbs related to *eat* leads them to conclude that there are, in fact, two different senses of *eat* in English. One of these, has roughly the sense of "ingest food in some manner" and the other, roughly, "eat a meal." *Nibble* and *dine* are respectively hyponyms of (i.e. subordinate to) these two different senses. Furthermore, they claim that only the "eat a meal" sense of *eat* permits indefinite objects to be omitted. For the manner-incorporating sense of eat, the direct argument must be overtly realized. They write:

> As Kegl and Fellbaum (1988) have argued, the presence of an obligatory adjunct always requires the presence of a direct argument. This is the reason that these manner verbs, which have absorbed the adjunct manner phrase, must have a d-structure direct argument that is overtly realized at s-structure.

Given this sense distinction, they argue, the behavior of verbs related to *eat* falls out of which sense they are subordinate to. Verbs that refer to manner of eating, like *gobble*, *gulp*, and *devour*, require overt direct objects. Verbs like *to breakfast*, *to dine*, and *to snack* are intransitive because they have incorporated the direct object, a kind of meal, into the verb itself. The "cross-category linking" of the paper's title refers to the parallel between the hyponyms of the canonical direct object *meal* — nouns *breakfast*, *picnic*, and so forth — and the corresponding denominals that are hyponyms of the verb.

This account is appealing because it offers a clean, semantically-driven account for an interesting range of data — not just *eat* and its relatives, but also *drink* (intransitive *to booze* vs. transitive *to guzzle*), *play* (intransitive *to drum* vs. transitive *to strum*), and the like. However, it does appear to have a central problem. Despite Fellbaum and Kegl's argument to the contrary, *eat* does permit the omission of indefinite objects even when they are not "understood as constituting some unit of food, i.e. a meal" (p. 97). For example, they argue that (154b) is not an appropriate answer to (154a) because nibbling all day cannot be construed as making up a meal.

(154) a. Have you eaten?
b. Yes, I've been nibbling all day.

However, I would argue that the oddness of (154b) arises less from this fact, and more from the question it customarily implies — usually something to the effect of "Do you want to go get lunch?" Changing the context makes it clear that the meal interpretation of the omitted object is in fact customary rather than obligatory: as an exchange between a doctor and a patient, example (154) is perfectly natural.[14] Similarly, if your friend utters (155a) at an amusement park just as you get on the roller coaster, (155b) seems to be a much more natural response than (155c).

(155) a. It's a bad idea to eat before doing this!

[14] I am indebted to Dan Hardt and Jamie Henderson for this observation. As it turns out, the very same observation can be found in (Rizzi, 1986, footnote 6).

b. Uh oh, I've been munching pretzels all day.
c. No problem, I've been munching pretzels all day.

Salvaging the argument would require that the food referred to in (155b) be construed as a meal, a claim that seems untenable. So, although it is true that verbs incorporating manner do not permit implicit indefinite objects, it is not at all clear that the rest of Fellbaum and Kegl's account can be made to work. In particular, cross-category linking between verbs and nouns in the taxonomy cannot be the whole story, since *eat* accepts implicit indefinite objects even when they are not construed as meals.

Where Kegl and Fellbaum attempt to trace the indefinite object alternation to a distinction between verb senses in the lexicon, Rice (1988) would like to do away with the traditional lexicon entirely. She writes:

> These differential usages [i.e., alternation between transitive and intransitive –PSR] do not arise from separate lexical entries for polysemous verbs. I will suggest, instead, that elements in the lexicon, if there is such a separate component of grammar, form natural categories that are subject to prototype effects and that many factors other than intrinsic meaning influence lexical insertion. (p. 202)

Later she asserts:

> In short, whether or not a transitive verb can omit its object . . . cannot possibly reside in the lexicon as a property of certain verbs because a lexicon with fixed lexical entries does not really exist. The lexicon is truly a convenient fiction . . . [Lexical] knowledge is best thought of as part of a dynamic interconnected network that can access sound, meaning, context, and speaker intent simultaneously. (p. 211)

It is not clear whether the strong form of this argument can be supported: the most suggestive evidence presented by Rice — cases where contextual influences license otherwise illicit omissions of the direct object — largely coincides with Kegl and Fellbaum's "discourse-conditioned" intransitivizations. One could claim that these involve traditional lexical entries together with pragmatically-controlled rules rather than a more holistic system.

Regardless of how the lexicon is construed, however, Rice posits a hypothesis about conditions for intransitivization that is particularly interesting with regard to the hypothesis pursued in this chapter. In addition to suggesting that a verb must have a "semantically neutral" or "basic-level" status in order to license an omitted object (echoing Kegl and Fellbaum's observations about manner incorporation), she comments that the omitted objects themselves tend to be interpreted as basic-level entities, illustrating with examples like (156):

(156) a. John smokes (*Marlboros/cigarettes/*smoking materials).
b. When he goes to Boston, John drives (*a Toyota/a car/*a vehicle).

A similar observation is made by Lehrer (1970), who distinguishes the "deletable object" from the selection restriction — for example, identifying the deletable object of *drive* as *car* and its selection restriction as VEHICLE.

However, this characterization seems to me to be too strong. Examples (157) and (158) make it clear that when these verbs omit their direct objects, the inferences are better described at a higher level.

Verb	Class
drink	⟨beverage⟩
drive	⟨vehicle⟩
eat	⟨food⟩
read	⟨writing⟩
smoke	⟨roll_of_tobacco⟩

Table 4.2: Some verbs and their associated classes of direct objects

(157) a. John smokes.
b. Ah! Therefore, John smokes cigarettes!
c. No — he smokes cigars.

(158) a. When he goes to Boston, John drives.
b. Ah! Therefore, when he goes to Boston, John drives a car.
c. No — he drives a van.

The confusion arises, I think, because it is odd to use the *label* for a superordinate category in contexts like (156) — a fact that arises more from conversational principles (Grice, 1975) than from inconsistency with the expectations of the verb. (For example, saying "John drank a beverage," implies that there is some reason for being less informative than is customary about what he drank.) What is important is not that the *words* "smoking materials" be natural in (156a), but rather that all the direct objects that *are* natural there be a member of that conceptual category. (The traditional superordinate categories may be too broad for this purpose — it might be odd to utter (156b) if what John drives to Boston is a snowplow, so perhaps the object category inferred from *drive* is more along the lines of "four-wheeled passenger vehicle.") So, although I would argue that basic-level categories are not the appropriate level of description, I agree that indefinite objects can only be omitted when the intended inferences about them can be captured at an appropriate "medium" level of abstraction.

Now, although the selectional preference criterion proposed in this chapter is expressed in terms of distributions over classes, rather than single categories, the measure of selectional association defined in Chapter 3 produces the kind of behavior that has just been described. Table 4.2 shows several object-drop verbs, each together with its single most strongly associated WordNet class.[15] In each case, the most strongly associated class fits intuitively as the "right level" of direct objects for the verb — a category that would seem to contain all the direct objects one could felicitously omit, while excluding most others. For example, *write* is more closely associated with the class of written materials than subordinates like ⟨essay⟩ or superordinates like ⟨communication⟩, *drive* is associated with vehicles rather than with cars or general conveyances (which would include trains and cargo ships), and *smoke* is associated with a class that includes the more specific cigarettes and cigars but not such non-tobacco narcotics as opium.

This behavior arises naturally from the definition of selectional association between a verb v and a class c (originally given in Chapter 3, equation 3.5):

$$\mathrm{A}(v,c) \quad = \quad \frac{1}{\eta}\mathrm{p}(c|v)\log\frac{\mathrm{p}(c|v)}{\mathrm{p}(c)}.$$

[15] These examples were constructed using co-occurrence statistics from the Brown Corpus.

Consider a single path through the taxonomy from a very specific class like ⟨fettuccine⟩ upward through ⟨thing⟩. As you move higher in the taxonomy, to ⟨pasta⟩, the superclass brings in additional objects of *eat*, such as *spaghetti* and *ravioli*, and still more objects of *eat* are brought in by continuing on to ⟨food⟩. As a result, moving upward in the taxonomy increases the conditional probability $p(c|v)$ and hence the selectional association with the verb. However, continuing further upward in the taxonomy, for instance from ⟨food⟩ up to ⟨substance⟩, brings in many words like *fuel* and *poison* that do *not* appear with *eat*. As a result, probability $p(c)$ increases without a corresponding increase in $p(c|v)$ and the score is driven back down. Thus, as this example illustrates, the measure of selectional association tends to prefer classes in the taxonomy that are general, but not too general.[16]

4.6.3 Summary

To summarize, a number of authors have investigated factors other than object inferability that might help determine whether a verb permits its objects to be omitted. These to a great extent revolve around features of meaning that may or may not be incorporated in to the verb, such as aspectual distinctions (whether or not a verb phrase can be interpreted as an activity, whether or not it is inherently completive) and manner. In several cases (Fellbaum and Kegl, 1989; Rice, 1988) underlying taxonomic relationships have been hypothesized among verbs and categories of their arguments in order to account for differences in behavior.

These semantic factors appear to account for much of the variability not captured by strength of selectional preference. For example, the strongest counterexamples to the selectional preference hypothesis — verbs that select strongly but cannot omit their objects — appear to be verbs like *catch*, *wear* and *say* that are difficult to interpret as activities when appearing intransitively. On inspection, at least, these aspectual distinctions, taken together with inferability on the basis of selectional preference, appear to provide a categorical distinction between verbs that do and do not permit implicit objects of the indefinite variety.

It makes sense that object realization depends on how much information is available. The proposal made here is consistent with intuitive notions about explicit mention, brevity, and clarity, such as Grice's (1975) maxims of Quantity.[17] In addition, the results, particularly in performance, are consistent with what we already know about other arguments. For instance, psycholinguistic experiments show that instruments (e.g. *John stabbed Bill with a knife*) are mentioned less frequently when typical for the given action (Brown and Dell, 1987), and "plausibility" of verb-argument relationships, often construed in probabilistic terms, is gaining increased attention in studies of on-line syntactic processing (Carlson and Tanenhaus, 1988; MacDonald, in revision; Mauner, Tanenhaus, and Carlson, 1992; Pearlmutter and MacDonald, 1993; Tanenhaus, Garnsey, and Boland, 1991; Trueswell, Tanenhaus, and Garnsey, 1993).

In contrast, although observations regarding aspect, manner, and taxonomy capture predictive generalizations about which verbs will and will not participate in implicit object alternations, they do so without providing an explanatory link between the relevant feature of meaning and the particular syntactic behavior it is connected to. Whatever deep relationship there may be between these factors and argument realization, an explanation of that connection will have to wait a better understanding of lexical semantics as a whole.[18]

[16] For a discussion of other probabilistic measures and their relationship to basic levels, see (Hanson, 1990).

[17] "Make your contribution as informative as is required (for the current purposes of the exchange)" and "Do not make your contribution more informative than is required."

[18] I would conjecture that Grimshaw's (1990) notion of *aspectual prominence* might be a useful place to start, since it provides a direct link between aspect and argument realization.

4.7 Thoughts on verb acquisition

4.7.1 Plausibility considerations

The results in this chapter show that two aspects of lexical representation — selectional constraints and optionality of an argument — can to a large extent be predicted on the basis of a corpus of text or transcribed speech, together with a simple taxonomic organization of noun concepts. Given the simplicity of the methods used here, a natural question to ask is whether the same ideas can contribute to a model of how lexical representations of verbs are acquired by children.

A first point in favor of such an approach is the relatively small number of assumptions that are required, and the psychological plausibility of those that are indispensible. To begin with the taxonomy, it is generally agreed that noun acquisition precedes verb acquisition (Nelson, 1973), and there is evidence to suggest that observation provides reliable evidence for learning how to map noun forms to noun concepts (Gillette and Gleitman, forthcoming). Furthermore, children at least as young as three years old can classify pictures of objects in the same manner as adults, at least for basic level categories such as TABLE and FISH, and sorting objects into superordinate categories such as FURNITURE and ANIMAL reaches adult competence by around the third grade (Rosch et al., 1976). So, although children's taxonomic criteria may not match those of adult taxonomies (much less the specifics of WordNet!), it is plausible to assume that they distinguish *some* form of category membership for observed instances — for example, permitting a red apple and a green apple (or a bunch of green beans, or a cookie) to be counted as instances of some class.[19]

The second element of the approach pursued here was observation of a sample of verb-argument co-occurrences, which would seem to require a procedure for identifying the argument. There is an accumulating body of evidence suggesting that children may be able to construct a skeletal parse on the basis of prosodic information (Gleitman et al., 1988; Kemler Nelson et al., 1989; Kemler Nelson, 1989; Lederer and Kelly, 1991), which could provide the basis for such a procedure, and it appears that the statistical methods demonstrated here are tolerant enough of noise to make do with very little parse information. For example, in the experiments done using data from CHILDES (and in earlier pilot experiments using the tagged Brown Corpus, done before the parses in the Penn Treebank became available), I found that a very unsophisticated object-finding procedure — little more than "select the first noun to the right" — yielded a noisy sample, but one for which the estimates of selectional preference and selectional association nonetheless yielded sensible results.

Given these assumptions — that the child can map noun forms to noun concepts, organizes noun concepts taxonomically, and can identify co-occurrences of verbs with noun arguments — the formalization of selectional preference proposed in Chapter 3 can be interpreted as a psycholinguistic model, and the algorithms involved in computing selectional preference from distributional evidence can be viewed as constituting a model of how such preferences are acquired. Furthermore, the central linguistic result of this chapter — that selectional preference is a predictor of object omissibility — represents a starting point for investigating how that aspect of lexical representation is acquired. The present study demonstrated that the predictive information is present in the child's input (represented using parental speech in CHILDES); the next necessary steps would be, first, to show that children attend to this information, and, second, to show that they actually make use of it.

[19] Care is needed to avoid circularity here, since in identifying class relationships to a verb like *eat* the relevant generalization of apples and cookies might turn out to be "things that you eat." Crucially, however, that characterization rests on the *concept* of eating, i.e. something like "things that you put in your mouth, chew, and swallow, etc." and not "things that co-occur with the word-form /eat/."

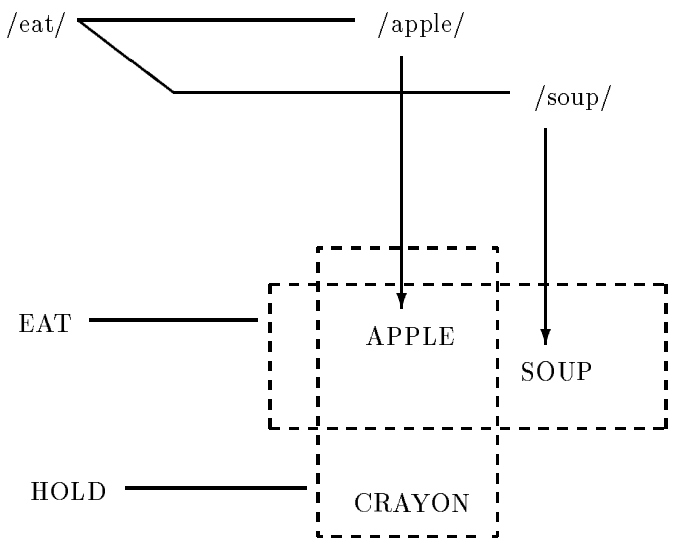


Figure 4.3: Schematic view of the verb mapping problem

4.7.2 Relation to bootstrapping

In addition to addressing the question of optional arguments, the model of selectional preference proposed here may be able to provide a necessary element in the general discussion of how verb meanings are acquired. Figure 4.3 illustrates one interpretation of the scenario confronting the language learner. At the top are word forms (represented simply as words within slashes) and at the bottom are word meanings or concepts (represented as words in uppercase). The child is assumed to possess a reliable mapping from familiar nouns to the concepts they represent, hence the arrows from noun forms to noun meanings. Dashed boxes represent higher-level concept classes — such as FOOD or SMALL SOLID OBJECT — which the child has acquired on the basis of observed similarities of form, function, or behavior. Finally, solid lines represent observed argument co-occurrences: at the top of the figure are observed co-occurrences between verb forms and nominal arguments (with respect to a particular argument position), and at the bottom are links between verb concepts and the classes containing noun concepts that have participated in the event (in a particular thematic role). Of course, syntactic arguments must somehow be mapped to thematic roles, but this is an issue that I will not attempt to address in detail. For the moment let us simply assume that the connection can be made via (universal) linking rules of the kind discussed in (Pinker, 1989).

Notably missing from the diagram is a connection between verb forms and verb meanings — filling in this link is one of the major problems the learner must solve.[20] In fact, there are really two distinct problems that need to be considered. The first concerns the identification of a particular verb with a "syntactically relevant semantic subclass" of verbs (Pinker, 1989, p. 107); that is, identifying aspects of a verb's meaning that concern its argument-taking properties and the kinds of syntactic alternations in which it can participate. The second problem has to do with identifying aspects of verb meaning that do *not* concern argument realization — for example, acquiring distinct meanings for verbs like *melt* and *boil*, or *slide* and *roll*, which are indistinguishable from the perspective of grammatical behavior. Pinker separates the two using the

[20] Assuming that the verb concepts already exist (corresponding to a process Pinker (1989)describes as "event-category labeling") is undoubtedly too simplistic; a more complete model would also have to provide for the generation of hypotheses about what event concepts to include in the lower left-hand corner of the figure.

evocative term "color-blind conservatism": the real-world color of a verb's argument will not be relevant for the first problem, though such cognitive distinctions may be crucial in solving the second one.

The first problem has been a source of some controversy, primarily concerning the sources of evidence available to the child language learner. One plausible hypothesis is that the grammatically relevant features of verb meaning can be learned by observing the co-occurrence of verb forms and events in the world; for example, an utterance containing /eat/ coinciding with the activity of eating something. Such a process is limited, however, by ambiguity in the interpretation of events: a child hearing the word form /pour/ as a glass is being filled with water from a pitcher will not know whether to associate /pour/ with pouring or with filling (or with holding the glass, tilting the pitcher, etc.). Pinker suggests that cross-situational analysis can be used to resolve the ambiguity:

> The ambiguity of what a verb means in a single situation, however, is eliminated by the behavior *across* situations. Though a given instance of filling a cup may be ambiguous between pouring and filling, *pour* but not *fill* will eventually be used when water is put in a glass up to the halfway mark, and *fill* but not *pour* will eventually be used when a glass is left on a windowsill in a rainstorm long enough to make it full. (p. 254)

Such a solution is intuitively appealing, and gains credibility from empirical evidence (see references cited in (Pinker, 1989)) and from computational experiments showing that cross-situational learning can be efficiently implemented and successfully applied in restricted computational settings (Siskind, 1992; Siskind, 1993b). However, Lila Gleitman and colleagues (Landau and Gleitman, 1985; Gleitman et al., 1988; Lederer, 1993) have argued that observation alone does not provide enough evidence to map verb forms to verb meanings, especially in cases where the described event is closed to observation (i.e. events described by mental verbs such as *know* and *believe*), or where an event supports hypothesized interpretations from multiple perspectives, as is the case with chasing events (which also contain fleeing), buying events (which also contain selling), and so forth. They take this to mean that learning via observation must be supported by additional constraints, and suggest that such constraints are provided by evidence about syntactic subcategorization available in the utterance. Evidence for this view comes from experiments showing that adults perform poorly on tasks of guessing the verb given an observed scene (Lederer, Gleitman, and Gleitman, forthcoming), that the syntactic context in which a novel verb is first presented influences children's interpretation of a scene described using that verb (Fisher et al., 1994), and that prosodic information in utterances can provide some phrase structure information and that children attend to such cues (Lederer and Kelly, 1991; Jusczyk et al., 1992).

An unresolved difficulty in these discussions is the absence of a precise characterization of what it means for a form of context to constrain hypotheses about verb meaning. For example, a useful indication of how constraining a context is, utilized by Lederer *et al.* (forthcoming) and Gillete and Gleitman (forthcoming), is the percentage of "correct" responses given by subjects — that is, responses matching the verb uttered during the scene.[21] Of course, simply measuring percent correct can be misleading: if the target verb was *call* and most of the incorrect responses to a scene were *talk*, surely the context should be judged more informative than a case where the error rate is the same but the incorrect responses are evenly distributed over a wide variety of other verbs. One way to avoid this difficulty, employed by the above-cited authors, is to consider not the absolute percent correct, but the frequency with which *incorrect* guesses are in the

[21] Scenes were presented as videotape clips with the sound turned off.

"semantic neighborhood" of the target verb. (See (Fisher, Gleitman, and Gleitman, 1991) for detailed discussion of how semantic neighborhoods are determined.)

The methods applied in this chapter suggest an alternative (or better, additional) measure of contextual predictiveness, namely the information-theoretic measure of relative entropy. That is, if Y is a random variable ranging over possible instances of a particular kind of context, and X ranges over events being predicted, then the predictability of that form of context can be measured by

$$\mathrm{D}(\mathrm{p}(x|y)||\mathrm{p}(x)) \quad = \quad \sum_{x,y} \mathrm{p}(x|y) \log \frac{\mathrm{p}(x|y)}{\mathrm{p}(x)}.$$

Applied to the situation just described, Y might range over videotaped scenes, and X over subjects' verb responses. As discussed in Chapter 3, the relative entropy between a prior distribution $\mathrm{p}(x)$ and a posterior distribution $\mathrm{p}(x|y)$ can be interpreted as how costly it would be, on average, to ignore the conditioning context y. This would seem to be precisely the kind of measure needed in order to evaluate the extent to which a form of context reduces the space of hypotheses about verb meanings.

Furthermore, using relative entropy makes it possible to consider different forms of context, and combinations of contexts, in a single unified framework where predictability is measured in bits of information. The sole requirement of such a framework is that it be possible to arrive at probability estimates for the contexts of interest. Where those are observed scenes, estimates can come from responses generated by human subjects, as in (Gillette and Gleitman, forthcoming; Fisher et al., 1994); where they are syntactic contexts such as subcategorization frames, large corpus resources such as the Penn Treebank are an alternative to experimental data.

Other forms of context can be considered, as well. In particular, the model in Figure 4.3 suggests that constraints on hypotheses about an unknown verb's meaning can be derived from knowledge about characteristic participants in events described by the verb — even if those events themselves are not witnessed by the learner. To consider an example, suppose that a learner is attempting to figure out what events the verb form /eat/ denotes. Many events involving eating also involve biting — a situation analogous to the overlap between *pour* and *fill* discussed earlier. Cross-situational analysis offers one solution: /bite/ will be uttered in some observed situations involving biting but not eating. However, argument/participant relationships reduce the reliance of such a strategy on the *observation* of scenes involving these actions: a learner who *hears* references to biting such as (159) will have received evidence that it is possible to bite things you don't typically eat, even if the correct mapping of /bite/ *is* is still a mystery.

(159) a. Betcha I bite your nose off if you keep screaming at me.
b. "Why do you want to bite that cat?" [Read from a book]

Such predictive information is known to be useful: Lederer *et al.* (forthcoming) have shown that in adult verb-guessing tasks, identifying the participants in an event provides useful predictive information, even if the roles are obscured by listing the participants in alphabetical order. Furthermore, Gropen (1992; 1993) has argued that children's categorization of objects in the world plays an important role in acquiring the meanings of polysemous verbs (i.e. verbs with multiple, related meanings, such as *spread*: *spreading pots and pans on the floor* vs. *spreading butter on toast*). Gropen points out that distinguishing such meanings is crucial for models of cross-situational analysis, since otherwise critical elements of meaning — e.g. forceful contact for the *spreading butter* sense of *spread* — will be true in some events and false in others, leading to their exclusion as core properties of the verb.

The model of selectional relationships proposed in Chapter 3 and applied here represents one component in an acquisition model that takes such information into account. Although such a model is not yet fully developed, most of the necessary pieces are in place: corpus resources like the Penn Treebank and CHILDES provide links between the word forms for verbs and their nominal arguments, and WordNet provides a mapping from noun forms to noun meanings and the superordinate concepts of which they are a part.

Another possible role for the present model of selectional relationships concerns elements of verb meaning that are not grammatically relevant, elements that are are usually set aside under the heading of cognitive distinctions by Pinker, Gleitman, Grimshaw, and others. For example, Pinker (p. 259) suggests a way in which a child might infer from (160) that the meaning of the verb encodes the manner of motion of the ball,

(160) The ball glipped into the room.

but is careful to distinguish this from inferring what the relevant manner of motion is. He writes:

> A grammar can "see" the difference between smearing and pouring, or between shouting and telling . . . because all of these distinctions can be stated in terms of the privileged semantic vocabulary that is available to it. . . . However, a grammar cannot "see" the difference between smearing and smudging, between shouting and whispering, between sliding and rolling, or between coating and covering. (p. 277)

In many cases, elements of meaning such as manner of motion involve properties of participants in the event — for example, rolling is associated with roundness. Therefore one could imagine that selectional relationships might help reduce the space of hypotheses about what particular dimension of meaning a verb encodes. Although the experiments done here are not particularly well suited to demonstrating this, since neither WordNet nor the corpora used make it easy to isolate the relevant properties, the data gathered so far provide one or two suggestive examples. Using co-occurrence frequencies from the CHILDES data, the class of objects most strongly selected for by *roll* is ⟨round_shape⟩; on inspection of the lexical co-occurrence data, it turns out that the two most frequent direct objects of *roll* are *ball* and *barrel*. Regarding whispering and shouting, objects of *shout* in the Brown Corpus include *abuse*, *bellicosity*, and *cry*; objects of *whisper* include *secret* and *explanation*. Browsing Roget's thesaurus leads quickly to the discovery that *secret* is a member of categories 528 (Concealment) and 522 (Interpretation), of which *whisper* and *explanation* respectively are members; *shout* and *cry* are both members of category 411 (Cry; vociferation). So, although observation of manner is still the most likely clue to the distinction between shouting and whispering, the nature of their objects (observed cross-situationally) also appears to provide some collateral evidence.

In sum, selectional relationships represent a source of information likely to be used by children in acquiring both semantically relevant and cognitively relevant aspects of verb meaning. Although ultimately psycholinguistic methods must bear the burden of demonstrating what forms of evidence children do and do not use, mathematical models of the kind proposed here serve an important purpose: they serve as testing ground for existing proposals, and provide insights that might not be available without first adopting a computational frame of mind.

Chapter 5

Semantic Classes and Syntactic Ambiguity

In this chapter, I investigate a second application of the model proposed in Chapter 3, exploring the use of the implemented model as a statistical method for resolving syntactic ambiguity in processing unconstrained text. I argue that a number of "every way ambiguous" constructions — in particular, prepositional phrase attachment, coordination, and nominal compounds — can be resolved by appealing to conceptual relationships such as selectional preference and semantic similarity, and that class-based, information-theoretic formalizations of these notions provide a practical way to do so.

5.1 Overview

One of the most pressing problems facing large-scale natural language applications is the explosion of analyses permitted by the grammar. Most parsers designed to cover large subsets of English produce an uncomfortably large number of analyses for even simple sentences; for example, using the Xtag system (Paroubek, Schabes, and Joshi, 1992), a sentence like (161a) will have on the order of ten to fifteen parses, including two analyses in which *Max meeting Ed* is interpreted as a nominal compound.[1]

(161) a. I saw the person that was annoyed by Max meeting Ed.
b. Ed identifies himself in terms of who he's met with today. He had a meeting with Max this morning so right now he's calling himself Max meeting Ed.

Given only the grammar, such an analysis has to be permitted in order to cover cases like (161b), which, contrived though it may be, illustrates a *possible* way in which *Max meeting Ed* could be interpreted as a compound of three nouns. Church and Patil (1982) point out that perfectly natural sentences can yield "hundreds, perhaps thousands" of parse trees. Furthermore, they show that the most serious ambiguity problems are associated with some of the most pervasive constructions in natural language, including coordinations, prepositional phrase attachment, and nominal compounds.

[1] This is true even when part-of-speech tagging is done first, since there is a strong tendency to prefer noun rather than verb tags for gerunds. I am grateful to Beth Anne Hockey and Christy Doran for pointing out this example.

Church and Patil suggest that until it has more useful constraints to resolve ambiguities, a parser can do little better than to efficiently record all the possible attachments and move on. Acquiring such constraints and using them is the subject of this chapter: I argue that syntactic choices can to a great extent be constrained by such semantic/conceptual relationships as lexical (selectional) preference and semantic similarity. I substantiate the claim by showing that class-based, information-theoretic formalizations of these relationships help in making accurate disambiguation decisions.

The chapter is structured as follows: I begin in Section 5.2 with a brief summary of the major families of strategies that have been proposed for resolving syntactic ambiguity. In Section 5.3, I consider a particular instance of syntactic ambiguity involving coordination and nominal compounds, developing a collection of disambiguation strategies that take advantage of different cues to the correct structure. The strategies are evaluated in a disambiguation experiment using training and test material from the Penn Treebank. In Section 5.4, I take a similar approach to the problem of prepositional phrase attachment ambiguity, evaluating the results in a computational experiment and comparing the method to other similar methods that have been proposed.[2] Finally, in Section 5.5, I briefly consider how techniques of this kind might be applied to the problem of disambiguating nominal compounds.

5.2 Parsing Preference Strategies

There is a long history of research on the use of parse preference strategies for resolving syntactic ambiguity, a literature too large to review here. The major approaches can briefly be summarized as follows:

- Structural strategies. The literature on parsing includes a number of strategies based on syntactic structure that have been argued to account for human performance — and human errors — on various forms of ambiguity. Among the most frequently cited are right association (Kimball, 1973), a preference for constituents to attach to the lowest node to the right in the partial parse tree, and minimal attachment (Frazier, 1979), a preference for choosing the attachment that would result in a parse tree with the fewest nodes. Crucially, such strategies depend only on configurations within parse trees, and not on extra-syntactic factors or even the identity of the lexical items involved.

- Referential strategies. Mark Steedman and colleagues (Crain and Steedman, 1985; Altmann and Steedman, 1988) have demonstrated effects of referential context on human performance in resolving syntactic ambiguities. They show that sentences usually inducing garden-path effects can often be interpreted naturally in contexts supporting the non-obvious reading, and that sentences usually interpreted without difficulty can be turned into garden paths by appropriate manipulations of context. For example, the context set up by (162a) helps override the the usual garden-path effect created by (162b).

(162) a. Two men on horseback decided to have a race. One man took his horse through the meadow, and the other chose a shorter route near the barn.
b. The horse raced past the barn fell.

- Lexical preference strategies. Lexical items often have typical kinds of phrases with which they associate — for example, the following dictionary entries illustrate a lexical association between the

[2] The work in this chapter on prepositional phrase attachment was done in collaboration with Marti Hearst.

verb *ask* and prepositional phrases involving *of* or *for*, between *put* and locative prepositional phrases, and between *win* and prepositional phrases involving *for*.

ask: To request of or for; solicit.

put: To place in a specified location; set.

win: To receive as a reward for performance.

Lexical preferences of this kind can be used in resolving ambiguous attachments, by choosing to attach a constituent at the site where preferences are best satisfied. Approaches along these lines have been suggested by (Ford, Bresnan, and Kaplan, 1982; Wilks, Huang, and Fass, 1985; Dahlgren and McDowell, 1986; Jensen and Binot, 1987).

An empirical study of these strategies by (Whittemore, Ferrara, and Brunner, 1990) shows that lexical preference plays a predominant role in predicting prepositional phrase attachments: they observe that in naturally-occurring data, lexical preferences (e.g., *arrive at*, *flight to*) provide more reliable attachment predictions than structural strategies, though referential success is also a contributing factor. Unfortunately, it seems clear that, outside of restricted domains, hand-encoding of preference rules will not suffice for unconstrained text. Information gleaned from dictionaries may provide a solution, but the problem of how to weight and combine preferences remains unsolved.

A more practical alternative may be the automated acquisition of lexical preference relationships using large corpora, a topic investigated by (Hindle and Rooth, 1991; Hindle and Rooth, 1993; Weischedel et al., 1989; Weischedel et al., 1991; Basili, Pazienza, and Velardi, 1991; Grishman and Sterling, 1992). Common to these acquisition methods is the use of a robust syntactic analyzer to obtain lexical co-occurrences of interest, together with some quantitative measure of association. Several of these investigations have also made use of relationships based on semantic word classes. Since many of these approaches have been applied to the problem of prepositional phrase attachment, I will defer a more detailed discussion until the end of Section 5.4.

5.3 Coordination

5.3.1 Cues to the correct analysis

Coordination is one of the most frequently occurring phenomena in natural text, and ambiguous coordinations are a common source of parsing difficulty. In this study, I investigated a particular subset of coordinations, noun phrase conjunctions of the form *noun1 and noun2 noun3*.[3] Examples of these include the following:

(163) a. a (bank and warehouse) guard
b. a (policeman) and (park guard)

(164) a. John is a (business and marketing) major
b. John is an (athlete) and (economics major)

Such structures admit two analyses, one in which *noun1* and *noun2* are the two heads being conjoined (163a) and one in which the conjoined heads are *noun1* and *noun3* (163b). A natural language system that

[3] All computational experiments in this chapter were performed using the earlier method for frequency estimation described in Appendix A.

analyzed (163b) according the structure in (163a) would be led to conclude that the noun phrase referred to someone who guards parks and policemen; similarly, analyzing (164a) according to the model in (164b) would lead one to conclude that John is a business. In each case, the "incorrect" analysis is one that is licensed by the grammar and perhaps even by knowledge about what is *possible* in the world, but constitutes at best a secondary reading.

As pointed out by Kurohashi and Nagao (1992), similarity of form and similarity of meaning are important cues to conjoinability. In English, similarity of form is to a great extent captured by agreement in number:

(165) a. several *business* and *university* groups
b. several *businesses* and university *groups*

Semantic similarity of the conjoined heads also appears to play an important role:

(166) a. a *television* and *radio* personality
b. a *psychologist* and sex *researcher*

Here, it is intuitively obvious that the correct structure is connected with the fact that televisions and radios have more in common than televisions and personalities, and that psychologists and researchers form a more natural category than psychologists and sex.

Finally, for this particular construction, the appropriateness of noun-noun modification for *noun1* and *noun3* is relevant:

(167) a. *mail* and securities *fraud*
b. *corn* and peanut *butter*

In general, phrases conjoining *noun1* and *noun2* are analyzed distributively, so that both are interpreted as modifying *noun3*. In (167b) the noun-noun compound *corn butter* is rather odd, providing a cue that that structure is inappropriate here.

5.3.2 Approximating the cues

Similarity of Form

In order to take advantage of the cues just described, it is necessary to approximate them in some computationally tractable way.

The first cue, similarity of form, is not difficult to approximate, since accurate reduction of nouns to their root form is well within the reach of automated methods. I reduced nouns to their root forms by doing a simple morphological analysis of suffixes in conjunction with lexical information from WordNet. Given *noun*, the reduction procedure had the following steps:

1. See if *noun* is a plural on WordNet's list of exceptional cases for noun pluralization (e.g. *oxen*, *ox*); if so, return the corresponding singular form.
2. For each suffix replacement rule *old* → *new*,
   (a) If *old* is a suffix of *noun*, strip it off and replace it with *new* to get *noun′*
   (b) If *noun′* is a noun in WordNet, halt and return it as the root form.
3. If no suffix-replacement rule applied, return *noun* itself as the root form.

Old	New	Example	
ss	ss	glass	glass
s	ϵ	bucks	buck
ses	s	glasses	glass
xes	x	boxes	box
zes	z	quizzes	quiz
ches	ch	matches	match
shes	sh	wishes	wish
ies	y	bodies	body
es	e	vases	vase
es	ϵ	tomatoes	tomato

Table 5.1: Suffix rules for reducing nouns to root form.

Table 5.1 lists the suffix replacement rules, (taken from WordNet 1.2 source code and included here with permission of the author); ϵ indicates the empty string. Given the algorithm just described, a noun can be considered plural if it differs from its root form, and singular otherwise. Naturally there are some unclear cases (e.g. *sheep* will unconditionally be labelled singular), but in general the simple suffix mappings, together with WordNet's large vocabulary and exceptions list, yield excellent results.

Similarity of Meaning

Many factors influence judgements of semantic similarity between two nouns; see, for example, (Cruse, 1986, Chapter 12) for an extensive discussion of considerations entering into judgements of synonymy. In addition, as discussed in Chapter 2, a great many researchers are investigating techniques for deriving measures of word similarity on the basis of distributional behavior. In the present investigation, I have opted to use taxonomic relationships in WordNet as the basis for an information-theoretic similarity measure. Like the formalization of selectional preference proposed in Chapter 3, this has the advantage of combining inductive, quantitative methods with an existing broad-coverage source of lexical knowledge. Furthermore, to the extent that relationships in WordNet can be given a formal semantic interpretation (see Chapter 2, Section 2.4.1), the similarity measure proposed here can be viewed as both mathematically and semantically well founded.

Before considering word similarity, it is helpful to consider the notion of *class* similarity in a taxonomy like WordNet. Intuitively, two noun classes in an IS-A taxonomy should be considered similar when there is a specific class that subsumes them both — if you have to travel very high in the taxonomy to find a class that subsumes both classes, in the extreme case all the way to the top, then they cannot have all that much in common. For example, ⟨`nickel`⟩ and ⟨`dime`⟩ are both immediately subsumed by ⟨`coin`⟩, whereas the most specific superclass that ⟨`nickel`⟩ and ⟨`mortgage`⟩ share is ⟨`possession`⟩.

The difficulty, of course, is how to measure "specific." Simply counting IS-A links in the taxonomy can be misleading, since a single link can represent a fine-grained distinction in one part of the taxonomy (e.g. ⟨`zebra`⟩ IS-A ⟨`equine`⟩) and a very large distinction elsewhere (e.g. ⟨`carcinogen`⟩ IS-A ⟨`substance`⟩). Counting other kinds of taxonomic links can be even more problematic; for example, (Morris and Hirst, 1991) point out that unbridled transitivity leads to spurious relatedness judgements through chains like {*cow,sheep,wool,scarf,boots,hat,snow*}.

Class c	$\log \frac{1}{\mathrm{p}(c)}$
⟨coin,3566679⟩	13.51
⟨coin,3566477⟩	12.52
⟨cash,3566144⟩	12.45
⟨currency,3565780⟩	11.69
⟨money,3565439⟩	11.27
⟨tender,3562119⟩	11.27
⟨medium_of_exchange,3561702⟩	11.21
⟨asset,3552852⟩	9.71
⟨possession,11572⟩	8.17

Table 5.2: Superclasses for $\langle nickel, 3567117\rangle$ and $\langle dime, 3567068\rangle$

An alternative to counting links is to consider the *information content* of a class as a way to measure its specificity. Information content of a class is defined in the standard way as negative the log likelihood, or $\log \frac{1}{\mathrm{p}(c)}$. The simplest way to compute similarity of two classes using this value would be to find the superclass that *maximizes* information content; that is, to define a similarity measure as follows:

$$\mathrm{sim}(c_1, c_2) \quad = \quad \max_{c_i} [\log \frac{1}{\mathrm{p}(c_i)}], \tag{5.1}$$

where $\{c_i\}$ is the set of classes dominating both c_1 and c_2, and the similarity is set to zero if that set is empty. For example, classes ⟨nickel⟩ (in the sense of a coin) and ⟨mortgage⟩ have only the superclass ⟨possession⟩ in common, with an information content of 8.17; classes ⟨nickel⟩ and ⟨dime⟩ have all the common superclasses listed in Table 5.2, the most specific of which yields a similarity score of 13.51.[4]

One natural way to measure word similarity is to consider all the classes to which a word belongs — that is, given two nouns n_1 and n_2, to compute their similarity as

$$\mathrm{sim}(n_1, n_2) \quad = \quad \max_{c_i} [\log \frac{1}{\mathrm{p}(c_i)}], \tag{5.2}$$

where $\{c_i\}$ is the set of all classes containing both n_1 and n_2.

Although there is not yet a standard way to evaluate computational measures of semantic similarity, one reasonable way to judge would seem to be agreement with human subjects on some relevant task. In a ratings task used by Miller and Charles (1991), subjects were given 30 pairs of nouns that were chosen to cover high, intermediate, and low levels of similarity (as determined using a previous study), and asked to rate "similarity of meaning" for each pair on a scale from 0 (no similarity) to 4 (perfect synonymy). In order to get a baseline against which to evaluate the performance of the information-theoretic similarity measure, I replicated Miller and Charles's experiment, giving ten subjects the same 30 noun pairs, five in a random order and the other five in the same random order reversed. The subjects were all computer science graduate students or postdocs, and the instructions were exactly the same as used by Miller and Charles, the main difference being that in this replication the subjects completed the questionnaire by electronic mail (though they were instructed to complete the whole thing in a single uninterrupted sitting).

[4] Class probabilities in this case were estimated using a sample of nouns from AP newswire.

n1	n2	sim(n1,n2)	class
tobacco	alcohol	10.84	⟨drug,1062813⟩
tobacco	sugar	7.76	⟨substance,5941⟩
tobacco	horse	11.85	⟨narcotic,1557422⟩

Table 5.3: Similarity with *tobacco* computed by maximizing information

The data from the experiment are given in Appendix C. On average, the correlation between the mean ratings in Miller and Charles's study and the the ratings of a subject in my replication was $r = 0.88$, with a standard deviation of 0.08 (inter-subject correlation in the replication, estimated using leaving-one-out resampling (Weiss and Kulikowski, 1991), was $r = .90$, stdev $= 0.07$). I evaluated the measure in equation (5.2) by treating it as if it were a subject in the same experiment. Owing to nouns missing from WordNet 1.2, it was possible to arrive at a rating for only 28 of the 30 pairs (93.3%); for that subset there was a correlation of $r = .77$ with Miller and Charles's means.[5] The average human subject correlation for those 28 stimuli was $r = 0.88$. As compared against that baseline, it seems clear that although there is certainly room for improvement, the information-based similarity measure is an entirely reasonable first approximation for human similarity judgements.

A problem with the similarity measure in equation (5.2) is that it sometimes produces spuriously high similarity measures for words on the basis of inappropriate word senses. For example, Table 5.3 shows the word similarity for several words with *tobacco*. *Tobacco* and *alcohol* are similar, both being drugs, and *tobacco* and *sugar* are similar, though less so, since both can be classified as substances. The problem arises, however, in the similarity rating for *tobacco* with *horse*: the word *horse* can be used as a slang term for *heroin*, and so the similarity rating is maximized when the two words are both categorized as narcotics. This is contrary to intuition.

The experimental evaluation using Miller and Charles's study suggests that cases like this are relatively rare. However, the example illustrates a more general concern: in measuring similarity between words, it is really the relationship among word *senses* that matters, and a similarity measure should be able to take this into account.

The most straightforward way to do so is to consider *all* classes to which both nouns belong rather than taking just the single maximally informative class. This suggests redefining class similarity as follows:

$$\text{sim}(c_1, c_2) \quad = \quad \sum_i \alpha(c_i)[\log \frac{1}{\text{p}(c_i)}], \tag{5.3}$$

where $\{c_i\}$ is the set of classes dominating both c_1 and c_2, as before, and $\sum_i \alpha(c_i) = 1$. This measure of similarity takes more information into account than the previous one: rather than relying on the single class with *maximum* information content, it allows *each* class to contribute information content according to the value of $\alpha(c_i)$. Intuitively, these α values measure relevance — for example, $\alpha(\langle\texttt{narcotic}\rangle)$ might be low in general usage but high in the context of a newspaper article about drug dealers.

[5] Class probabilities in this experiment were estimated using noun frequencies in the Brown corpus.

Equation (5.3) leaves α to be specified externally — by a word sense disambiguation algorithm, perhaps, or by whatever other means are available. Notice that if $\alpha(c_i)$ is fixed at 1 for the single c_i maximizing $\log \frac{1}{\mathrm{p}(c_i)}$ and at 0 for $c_j, j \neq i$, then equation (5.3) simply reduces to the "global" measure in equation (5.1).

Appropriateness of noun-noun modification

Judging the "goodness of fit" between a modifier (n_m) and a head (n_h) is not unlike judging the goodness of fit between a verb and its object — in both cases, the judgement can be made in terms of selectional association between a selecting word and a class of nouns being selected for. In the case of nominal modification, since both head and modifier are nouns, there are two selectional relationships that can be considered: selection of the modifier for the head, and selection of the head for the modifier. That is, for a particular class c_h containing the head, we can define

$$\mathrm{A}(n_m \rightarrow c_h) \quad = \quad \frac{\mathrm{p}(c_h|n_m) \log \frac{\mathrm{p}(c_h|n_m)}{\mathrm{p}(c_h)}}{\sum_c \mathrm{p}(c|n_m) \log \frac{\mathrm{p}(c|n_m)}{\mathrm{p}(c)}}. \tag{5.4}$$

Correspondingly, for a particular class c_m containing the nominal modifier,

$$\mathrm{A}(c_m \leftarrow n_h) \quad = \quad \frac{\mathrm{p}(c_m|n_h) \log \frac{\mathrm{p}(c_m|n_h)}{\mathrm{p}(c_m)}}{\sum_c \mathrm{p}(c|n_h) \log \frac{\mathrm{p}(c|n_h)}{\mathrm{p}(c)}}. \tag{5.5}$$

The "goodness" of a particular noun-noun compound $n_m\ n_h$ can be evaluated by examining the strength of selectional association between modifier and head, and vice versa. The simplest way to do so is to see whether in either case selectional association exceeds a threshold, τ. By inspection, $\tau = 2.0$ seems to be a reasonable value for this threshold.

Consider, for example, the ambiguous coordinations in (168):

(168) a. They bought a new computer and telephone network for the office.
b. They bought a new computer and water cooler for the office.

Selectional association indicates that *computer network* is a reasonable nominal compound, since $\mathrm{A}(\langle \texttt{computer,1277690} \rangle \leftarrow$ *network*$) = 2.2$, $\mathrm{A}($*computer* $\rightarrow \langle \texttt{system,278118} \rangle) = 2.56$, *network* $\in \langle \texttt{system,278118} \rangle$, but that *computer cooler* is not, since $\mathrm{A}(c \leftarrow$ *cooler*$) < 2.0$ for all classes c containing *computer*; $\mathrm{A}($*computer* $\rightarrow c) < 2.0$ for all c containing *cooler*.

It is worth repeating the observation from Chapter 3 that selectional association between modifiers and heads accomplishes a limited form of word-sense disambiguation. For example, consider the constraints that the modifier places on the head in the compound *newspaper article*: the selectional association $\mathrm{A}($*newspaper* $\rightarrow \langle \texttt{news,2298043} \rangle) = 2.08$, whereas $\mathrm{A}($*newspaper* $\rightarrow \langle \texttt{function\_word,2216900} \rangle) = 0.55$, in effect showing that in the context of being modified by *newspaper*, the "news" sense of *article* is more relevant than its grammatical sense. Head-modifier constraints behave the same way: $\mathrm{A}(\langle \texttt{material,3886012} \rangle \leftarrow$ *article*$) = 0.95$ and $\mathrm{A}(\langle \texttt{press,2200204} \rangle \leftarrow$ *article*$) = 2.27$, where $\langle \texttt{press,2200204} \rangle$ is the WordNet class glossed as "printed matter in the form of newspapers or magazines." Thus as a nominal modifier for *article*, the word *newspaper* is better construed as printed matter than as a kind of physical material.

5.3.3 Experiment 1

I investigated the roles of the various cues to coordination by conducting a disambiguation experiment using the definitions just discussed. Two sets of 100 noun phrases of the form [NP *noun1 and noun2 noun3*] were extracted from the *Wall Street Journal* (WSJ) corpus in the Penn Treebank and disambiguated by hand, with one set to be used for development and the other for testing.[6] A set of simple transformations was applied to all WSJ data, including the mapping of all proper names to the token *someone*, the expansion of month abbreviations, and the reduction of all nouns to their root forms.

Similarity of form was determined as described above, and similarity of meaning was determined "globally" as in equation (5.2) using noun class probabilities estimated from a sample of approximately 800,000 noun occurrences in Associated Press newswire stories.[7] For the purpose of determining semantic similarity, nouns not in WordNet were treated as instances of the class $\langle \texttt{thing} \rangle$. Appropriateness of noun-noun modification was determined as described above in equations (5.4) and (5.5), with co-occurrence frequencies calculated using a sample of approximately 15,000 noun-noun compounds extracted from the WSJ corpus. (This sample did not include the test data.)

Each of the three sources of information — form similarity, meaning similarity, and modification relationships — was used alone as a disambiguation strategy, as follows:

- Form:
  - If noun1 and noun2 match in number
    and noun1 and noun3 do not
    then conjoin noun1 and noun2;
  - if noun1 and noun3 match in number
    and noun1 and noun2 do not
    then conjoin noun1 and noun3;
  - otherwise remain undecided.
- Meaning:
  - If sim(noun1,noun2) $>$ sim(noun1,noun3)
    then conjoin noun1 and noun2;
  - if sim(noun1,noun3) $>$ sim(noun1,noun2)
    then conjoin noun1 and noun3;
  - otherwise remain undecided.
- Modification:
  - If A$(noun1 \rightarrow noun3) > \tau$, a threshold, *or*
    if A$(noun1 \leftarrow noun3) > \tau$,
    then conjoin noun1 and noun3;
  - If A$(noun1 \rightarrow noun3) < \sigma$ *and* A$(noun1 \leftarrow noun3) < \sigma$
    then conjoin noun1 and noun2;

[6] Hand disambiguation was necessary because the Penn Treebank does not encode NP-internal structure. These phrases were disambiguated using the full sentence in which they occurred, plus the previous and following sentence, as context.

[7] I am grateful to Donald Hindle for making these data available.

– otherwise remain undecided.[8]

In addition, I investigated several methods for combining the three sources of information. These included:

- "Backing off"
  Use the form strategy if not undecided;
  otherwise use the modification strategy if not undecided;
  otherwise use the meaning strategy if not undecided;
  otherwise remain undecided.

- Voting
  Tally the votes of the three individual strategies;
  use the majority, if there is one;
  otherwise remain undecided.

- Regression
  Represent training instances as vectors of attributes;
  represent the two bracketings as -1 and 1;
  perform a linear regression;
  classify test instances using the regression equation.

- Decision tree
  Represent training instances as vectors of attributes;
  represent the two bracketings as classes;
  construct a decision tree classifier;
  classify test instances using the tree.

The training set contained a bias in favor of conjoining *noun1* and *noun2*, so a structural "default" strategy — always choosing that bracketing — was used as a baseline. The results were as follows:

STRATEGY	COVERAGE (%)	ACCURACY (%)
Default	100.0	66.0
Form	53.0	90.6
Modification	75.0	69.3
Meaning	66.0	71.2
Backing off	95.0	81.1
Voting	89.0	78.7
Regression	100.0	79.0
ID3 Tree	100.0	80.0

Not surprisingly, the individual strategies perform reasonably well on the instances they can classify, but coverage is poor; the strategy based on similarity of form is highly accurate, but arrives at an answer only half the time. Of the combined strategies, the "backing off" approach succeeds in answering 95% of the time and achieving 81.1% accuracy — a reduction of 44.4% in the baseline error rate. Although this confirms that there is useful predictive power in the meaning and modification strategies, a caveat is in order: this

[8] Thresholds τ and σ were fixed before evaluating the test data.

reduction in error may not be a thoroughly convincing demonstration of effectiveness since, on the basis of the above numbers, backing off from form to the default strategy could theoretically be expected to have an 84% accuracy at 100% coverage. This concern was addressed by the experiment that follows.

5.3.4 Experiment 2

In order to better evaluate bottom-line performance, I investigated the disambiguation of more complex coordinations of the form [NP *noun1 noun2 and noun3 noun4*], which permit five possible bracketings:

(169) a. freshman ((business and marketing) major)
b. (food (handling and storage)) procedures
c. ((mail fraud) and bribery) charges
d. Clorets (gum and (breath mints))
e. (baby food) and (puppy chow)

These bracketings comprise two groups, those in which the conjoined heads *noun2* and *noun3* (a–c) and those in which the conjoined heads are *noun2* and *noun4* (d–e). Rather than tackling the five-way disambiguation problem immediately, I used an experimental task of classifying a noun phrase as belonging to one of these two groups, thus providing a closer parallel to Experiment 1.

I examined three classification strategies. First, I used the form-based strategy described above. Second, as before, I used a strategy based on semantic similarity; this time, however, selectional association was used to determine the α_i in equation (5.3), incorporating modifier-head relationships into the semantic similarity strategy. That is, given *noun1 noun2 and noun3 noun4*, the similarity of *noun2* and *noun3* was calculated as

$$\text{sim}(noun2, noun3) \quad = \quad \sum_i \alpha(c_i)[\log \frac{1}{\text{p}(c_i)}], \tag{5.6}$$

where $\{c_i\}$ is the set of all classes containing both *noun2* and *noun3*, and

$$\alpha(c_i) \quad = \quad \frac{A(c_i \leftarrow \textit{noun4})}{\sum_j A(c_j \leftarrow \textit{noun4})}. \tag{5.7}$$

Notice that this similarity calculation takes advantage of more information than the "global" similarity. Intuitively, equation (5.6) computes the similarity of the two nouns *in the context of being modifiers to noun4*, where the role of context is determined in equation (5.7) on the basis of selectional association.

Similarly, in calculating the similarity between *noun2* and *noun4*, it was possible to take advantage of the additional information provided by *noun3*:

$$\text{sim}(noun2, noun4) \quad = \quad \sum_i \alpha(c_i)[\log \frac{1}{\text{p}(c_i)}], \tag{5.8}$$

$$\alpha(c_i) \quad = \quad \frac{A(\textit{noun3} \rightarrow c_i)}{\sum_j A(\textit{noun3} \rightarrow c_j)}. \tag{5.9}$$

Here, the similarity of *noun2* and *noun4* must be considered in light of the fact that *noun4* is modified by *noun3*.

As a third strategy, I used "backing off" (from form similarity to semantic similarity) to combine the two individual strategies. As before, one set of items was used for development, and another set (89 items) was set aside for testing. As a baseline, results were evaluated against a simple default strategy of always choosing the group that was more common in the development set.

STRATEGY	COVERAGE (%)	ACCURACY (%)
Default	100.0	44.9
Form	40.4	80.6
Meaning	69.7	77.4
Backing off	85.4	81.6

In this case, the default strategy defined using the development set was misleading, leading to worse than chance accuracy. However, even if default choices were made using the bias found in the test set, accuracy would be only 55.1%. The results in the above table make it clear that the strategies using form and meaning are far more accurate, and that combining them leads to coverage and accuracy that would not have been possible using similarity of form alone.

The pattern of results in these two experiments demonstrates a significant reduction in syntactic misanalyses for this construction as compared to the simple baseline, and it confirms that form, meaning, and modification relationships all play a role in disambiguation. In addition, these results confirm the practical effectiveness of the proposed definitions of selectional preference and semantic similarity.

5.4 Prepositional Phrase Attachment

Prepositional phrase attachment is the paradigm case for discussions of syntactic ambiguity. Examples (170,171, 172), from (Church and Patil, 1982), illustrate both the explosion of analyses as the number of prepositional phrases grows and the need for extra-syntactic constraints to select among them.

(170) Put the block [on the table].

(171) a. Put the block [in the box on the table].
b. Put [the block in the box] on the table.

(172) a. Put the block [[in the box on the table] in the kitchen].
b. Put the block [in the box [on the table] in the kitchen]].
c. Put [[the block in the box] on the table] in the kitchen.
d. Put [the block [in the box on the table]] in the kitchen.
e. Put [the block in the box] [on the table in the kitchen].

Resolving the ambiguity in this example seems to require information and inferences about the situation, since nothing about the lexical items provides constraints on their possible relationships. If all attachment ambiguities required that level of knowledge, life would indeed be difficult — Hindle and Rooth (1993, p. 103) comment:

> [One] recent proposal suggests that resolving attachment ambiguity requires the construction of a discourse model in which the entities referred to in a text are represented and reasoned about . . . We take this argument to show that reasoning essentially involving reference in a discourse model is implicated in resolving attachment ambiguities in a certain class of cases. If this phenomenon is typical, there is little hope in the near term for building computational models capable of resolving such ambiguities in unrestricted text.

Fortunately, however, lexical relationships can provide a great deal of guidance in attachment decisions, even in the absence of discourse context. As discussed earlier, a study by Whittemore *et al.* (1990) found lexical preferences to be a strong predictor of attachment, as illustrated in (173) (their (4)):

(173) a. What is the round trip fare for Aer Lingus and for British Airlines from JFK on August 30 to Dublin returning September 21?
b. What is the round trip [fare ... [from JFK] [to Dublin] ...]

The example shows that knowledge about preferred prepositions — in this domain, relationships like *fare from X* and *fare to Y* — suffices to predict the correct attachments solely on the basis of the lexical items involved.

5.4.1 Lexical association

As mentioned earlier, a critical obstacle to using this kind of information on a large scale is the difficulty in acquiring a collection of lexical preference relationships. Hindle and Rooth (1991; 1993) propose to overcome this obstacle using corpus-based lexical co-occurrence statistics.

The problem setting adopted by Hindle and Rooth is a sub-case of the general attachment problem, involving a choice between just two attachment sites. An "instance" of ambiguous prepositional phrase attachment in this setting consists of a verb, its direct object, a preposition, and the object of the preposition. Furthermore, only the heads of the respective phrases are considered; so, for example, the ambiguous attachment in (170) would be construed as the 4-tuple (*put*,*block*,*on*,*table*). Its elements will be called v, n1, p, and n2, respectively.

The attachment strategy is based on an assessment of how likely the preposition is, given each potential attachment site; that is, a comparison of the values p$(p|n1)$ and p$(p|v)$. For (170), one would expect p$(on|put)$ to be greater than p$(on|block)$, reflecting the intuition that *put X on Y* is more plausible as a verb phrase than *block on Z* is as a noun phrase.

Hindle and Rooth extracted their training data from a corpus of Associated Press news stories. A robust parser (Hindle, 1983) was used to construct a table in which each row contains the head noun of each noun phrase, the preceding verb (if the noun phrase was the verb's direct object), and the following preposition, if any occurred. Attachment decisions for the training data in the table were then made using a heuristic procedure — for example, given *spare it from*, the procedure would count this row as an instance of *spare from* rather than *it from*, since a prepositional phrase cannot be attached to a pronoun. Not all the data can be assigned with such certainty: ambiguous cases in the training data were handled either by using statistics collected from the unambiguous cases, by splitting the attachment between the noun and the verb, or by defaulting to attachment to the noun.

Given an instance of ambiguous prepositional phrase attachment from the test set, Hindle and Rooth used a statistical test to assess the direction and significance of the difference between p$(p|n1)$ and p$(p|v)$, a procedure they call *lexical association*. In (Hindle and Rooth, 1991) they used the t-score (Church et al., 1991) as their test, and in (Hindle and Rooth, 1993) they shifted to a log likelihood ratio. In both the earlier and later versions of the work, the value produced by their test is positive, zero, or negative according to whether p$(p|v)$ is greater, equal to, or less than p$(p|n1)$, respectively, and its magnitude indicates a level of confidence in the significance of this difference.

On a set of test sentences held out from the training data, the lexical association procedure used in (Hindle and Rooth, 1991) (t-score) made the correct attachment 78.3% of the time. For choices with a high level of

confidence (magnitude of t greater than 2.1, about 70% of the time), correct attachments were made 84.5% of the time. Using the log likelihood ratio in (Hindle and Rooth, 1993), they obtained a correct decision 79.7% of the time; for high confidence choices (log likelihood ratio greater than 2.0) they obtained 88.7% accuracy at 70.6% coverage.

5.4.2 Prepositional objects

The lexical association strategy performs quite well, despite the fact that the object of the preposition is ignored. However, Hindle and Rooth note that neglecting this information can hurt in some cases. For instance, the lexical association strategy is presented with exactly the same information in (174a) and (174b), and is therefore unable to distinguish them.

(174) a. Britain reopened its embassy in December.
b. Britain reopened its embassy in Teheran.

Furthermore, (Hearst and Church, in preparation) have conducted a pilot study in which human subjects are asked to guess prepositional phrase attachments despite the omission of the direct object, the object of the preposition, or both. The results of this study, though preliminary, suggest that the object of the preposition contributes an amount of information comparable to that contributed by the direct object; more important, for some prepositions, the object of the preposition appears to be *more* informative.

Thus, there appears to be good reason to incorporate the object of the preposition in lexical association calculations. The difficulty, of course, is that the data are far too sparse to permit the most obvious extension. Attempts to simply compare p$(p, n2|n1)$ against p$(p, n2|v)$ using the t-score fail dismally, and there is no reason to think the log likelihood ratio would fare any better.[9]

We are faced with a well-known tradeoff: increasing the number of words attended to by a statistical language model will in general tend to increase its accuracy, but doing so increases the number of probabilities to be estimated, leading to the need for larger (and often impractically larger) sets of training data in order to obtain accurate estimates. One option is simply to pay attention to fewer words, as do Hindle and Rooth. Another possibility, however, is to reduce the number of parameters by grouping words into equivalence classes, as discussed, for example, by (Brown et al., 1990). Figure 5.1 illustrates the intuition behind such an approach. Resolving the attachment ambiguity in the figure, it does not really matter that the particular city is Dallas — it could just as easily be any other city, and the attachment decision would be the same. Similarly, the specific word *staff* is not crucial since a number of other related words would produce exactly the same result. These intuitions suggest that the use of classes may be more than a useful engineering solution to the problem of data sparseness — the relevant relationships really do seem to obtain not at the lexical level, but at the level of classes or concepts.

5.4.3 Conceptual association

The preceding discussion suggests that class-based statistical relationships of the kind exploited in Section 5.3 may also be useful for prepositional phrase attachment. One might call such a proposal *conceptual association*: calculating a measure of association using the classes to which the direct object and object of the preposition belong, and selecting the attachment site for which the evidence of association is strongest.

[9] I attempted this experiment using expected likelihood estimates, as in (Hindle and Rooth, 1991), with data extracted from the Penn Treebank as described below.

a. He flies two personal secretaries in from Little Rock to augment his staff in Dallas.

b. ...augment his { personnel, team, crew, group, organization, people, ⋮ } in { Pittsburgh, Berkeley, Newark, Columbus, ⋮ }.

Figure 5.1: Noun classes in prepositional phrase attachment

The use of classes introduces two sources of ambiguity. The first, shared by lexical association, is word sense ambiguity: just as lexically-based methods conflate multiple senses of a word into the count of a single token, here each word may be mapped to many different classes in the WordNet taxonomy. Second, even for a single sense, a word may be classified at many levels of abstraction — for example, even interpreted solely as a physical object (rather than a monetary unit), *penny* may be categorized as a ⟨`coin,3566679`⟩, ⟨`cash,3566144`⟩, ⟨`money,3565439`⟩, and so forth on up to ⟨`possession,11572`⟩.

In the algorithm that follows, the simplest possible approach to these ambiguities was taken: *each* classification of the nouns is considered as a source of evidence about association, and these sources of evidence combined to reach a single attachment decision.

Algorithm 1. Given $(v, n1, p, n2)$,

1. Let C1 = $\{c \mid n1 \in \text{words}(c)\}$
   Let C2 = $\{c \mid n2 \in \text{words}(c)\} = \{c_{2,1}, \ldots, c_{2,N}\}$

2. For i from 1 to N,

$$c_{1,i} = \underset{c \in \text{C1}}{\text{argmax}}\ \text{I}(c; p, c_{2,i})$$

3. For i from 1 to N,

$$I_i^v = \text{I}(v; p, c_{2,i})$$
$$S_i^v = \text{freq}(v, p, c_{2,i})\ I_i^v$$

$$I_i^n = \text{I}(c_{1,i}; p, c_{2,i})$$
$$S_i^n = \text{freq}(c_{1,i}, p, c_{2,i})\ I_i^n$$

4. Compute a paired samples t-test for a difference of the means of S^n and S^v. Let "confidence" be the significance of the test with $N - 1$ degrees of freedom.

5. Select attachment to n1 or v according to whether t is positive or negative, respectively.

Step 1 of the algorithm establishes the range of possible classifications for n1 and n2. For example, if the algorithm is trying to disambiguate the example in Figure (5.1), the verb attachment *augment in Dallas* can be construed according to the following various classifications of *Dallas*:

v	p	c2
augment	in	Dallas
		⟨Dallas⟩
		⟨urban_area⟩
		⟨region⟩
		⟨geographical_area⟩
		⟨city⟩
		⟨location⟩

In step 2, each candidate classification for n2 is held fixed, and a classification for n1 is chosen that maximizes the association (as measured by mutual information) between the noun-attachment site and the prepositional phrase. In effect, this answers the question, "If we were to categorize n2 in this way, what would be the best class to use for n1?"

c1	p	c2
staff	in	Dallas
⋮		
⟨social_group⟩	in	⟨region⟩
⟨faculty⟩		
⟨implement⟩		
⟨symbol⟩		
⟨body⟩		
⟨musical_notation⟩		
⟨personnel⟩		
⟨assemblage⟩		
⋮		

For example, if *Dallas* is categorized in class ⟨region⟩, then, of all the classes to which *staff* belongs, the one maximizing mutual information would be chosen — in this case, ⟨personnel⟩. This is done for each classification of n2, yielding N different class-based interpretations for (n1,p,n2):

c1	p	c2
staff	in	Dallas
⟨gathering⟩	in	⟨dallas⟩
⟨people⟩	in	⟨urban_area⟩
⟨personnel⟩	in	⟨region⟩
⟨personnel⟩	in	⟨geographical_area⟩
⟨people⟩	in	⟨city⟩
⟨personnel⟩	in	⟨location⟩

Steps 1 and 2 result in N different classifications of the 4-tuple according to WordNet classes. In Step 3, each of these is given a score evaluating attachment to the verb and a score evaluating attachment to the noun, producing a table like the following:[10]

Classification			S^v	S^n
augment	⟨gathering⟩	in ⟨dallas⟩	45.54	38.18
augment	⟨people⟩	in ⟨urban_area⟩	28.46	1200.21
augment	⟨personnel⟩	in ⟨region⟩	23.38	314.62
augment	⟨personnel⟩	in ⟨geo._area⟩	26.80	106.05
augment	⟨people⟩	in ⟨city⟩	28.61	1161.22
augment	⟨personnel⟩	in ⟨location⟩	22.83	320.85

In the absence of sense disambiguation, there is no way to tell which classification of the nouns is most appropriate: the best one can do under the circumstances is to ask whether one attachment tends to score higher than the other across the different classifications. Step 4 implements an extremely brute-force way of asking this question: a t-test for the difference of the means is performed, treating S^n and S^v as paired samples (see, e.g., (Woods, Fletcher, and Hughes, 1986)). In step 5 the resulting value of t determines the choice of attachment site, as well as an estimate of how significant the difference is between the two alternatives. (For this example, $t(3) = 3.57, p < 0.05$, yielding the correct choice of attachment.)

In addition to evaluating the performance of the conceptual association strategy in isolation, it is natural to combine the predictions of the lexical and conceptual association strategies to make a single prediction. Although better-motivated strategies for combining the predictions of multiple models do exist (e.g. (Jelinek and Mercer, 1980; Katz, 1987)), a simpler "backing off" style procedure has been pursued here. The central idea behind backing off is to use the model making the most accurate predictions first, even if its coverage is poor; in cases where the more accurate model fails to apply (or makes a low-confidence decision), the less accurate model can be used.

Algorithm 2. Given $(v, n1, p, n2)$,

1. Calculate an attachment decision using lexical association (t-score).
2. If confident ($|t| > 2.1$), use this decision.
3. Otherwise, calculate an attachment using conceptual association

[10] The score used here is quite similar, though not identical, to selectional association of the verb and noun for the prepositional phrase. That would be computed as

$$A(c1; p, c2) = \frac{\text{freq}(p, c2|c1)\text{I}(c1; p, c2)}{\sum_{p', c2'} \text{freq}(p', c2'|c1)\text{I}(c1; p', c2')},$$

whereas this score weights mutual information by the joint rather than the conditional probability (frequency) and does not normalize:

$$S^n = \text{freq}(c1, p, c2)\text{I}(c1; p, c2).$$

I used weighted mutual information primarily because it was more straightforward to implement, and because it can be viewed simply as using mutual information together with an added factor to give more weight to higher-frequency (and hence more reliable) co-occurrences. However, using selectional association would be more consistent with the rest of the dissertation and is a topic for future study.

4. If confident ($p < 0.1$), use this decision;

5. Otherwise,

   (a) If confident decisions only are required, make no decision.

   (b) If a decision is needed in every case, use the choice made in step 3.

(Note: in earlier work (Resnik and Hearst, 1993; Resnik, 1993), backing off was done from conceptual association to lexical association, rather than vice versa, by analogy with backing off from bigrams to trigrams. The poorer results reported in that work reflect the fact that, although conceptual association gains information by paying attention to the object of the preposition, it sacrifices accuracy by abstracting to classes.)

5.4.4 Experimental results

Quantitative evaluation

An experiment was conducted to evaluate the performance of the lexical association, conceptual association, and backing off strategies. The corpus used was a collection of parses from articles in the 1988–89 *Wall Street Journal*, found as part of the Penn Treebank. This corpus is an order of magnitude smaller than the one used by Hindle and Rooth in their experiments, but it provides considerably less noisy data, since parse trees have been produced automatically by the Fidditch parser (Hindle, 1983) and then corrected by hand.

A test set of 201 ambiguous prepositional phrase attachment instances was set aside. After acquiring attachment choices on these instances from a separate judge (who used the full sentence context in each case), the test set was reduced by eliminating sentences for which the separate judge disagreed with the Treebank, leaving a test set of 174 instances.[11]

Lexical counts for relevant prepositional phrase attachments (v,p,n2 and n1,p,n2) were extracted from the parse trees in the corpus; in addition, by analogy with Hindle and Rooth's training procedure, instances of verbs and nouns that did not have a prepositional phrase attached were counted as occurring with the "null prepositional phrase." A set of clean-up steps included reducing verbs and nouns to their root forms, mapping to lowercase, substituting the word *someone* for nouns not in WordNet that were part-of-speech-tagged as proper names, substituting the word *amount* for the token *%* (this appeared as a head noun in phrases such as *rose 10 %*), and expanding month abbreviations such as *Jan.* to the full month name.

When each strategy was required to make a choice, regardless of level of confidence, the results were as follows:

STRATEGY	ACCURACY (%)	COVERAGE (%)
LA	81.6	100.0
CA	79.3	100.0
COMBINED	83.9	100.0

When each strategy was permitted to make a choice only when confident, the results were as follows:

[11] Of the 348 nouns appearing as part of the test set, 12 were not covered by WordNet ; these were classified by default as members of the WordNet class ⟨`entity,2383`⟩.

STRATEGY	ACCURACY (%)	COVERAGE (%)
LA	92.3	52.3
CA	83.9	67.8
BACKING OFF	88.5	79.9

The results at 100% coverage suggest that conceptual association alone is not taking full advantage of the additional information it has as compared to lexical association; reasons for this are discussed below. When levels of confidence are taken into account, it is evident that conceptual association increases coverage (by about 30%) at some cost to accuracy (about 9%). Combining the two strategies appears to be a successful way to offset the losses in this tradeoff: the loss of accuracy (about 4%) is a relatively small price to pay for increasing coverage by more than half (53%).

Although it is difficult to make comparisons between experiments using different sets of training and test data, it is worth noting that the performance of lexical association in this experiment is comparable with the recent results reported in (Hindle and Rooth, 1993) — in particular, they report 92.3% accuracy with 54.3% of the test cases covered. On their precision-recall curve, their coverage at 88.7% accuracy is 70.6%.

Qualitative evaluation

The quantitative results reported here could be considered equivocal: using class-based statistics appears to provide significant improvements in the coverage/accuracy tradeoff, but only a marginal increase at 100% coverage, at the cost of a fair amount of extra machinery. There are several reasons this may be the case.

The most obvious problem with conceptual association as implemented here is the cavalier way it handles multiple class membership. Although the class for n1 is chosen with attention to the prepositional phrase (step 2 of Algorithm 1), all possible classes to which n2 might belong are considered, and worse, weighted together equally using the paired t-test. As a result, although abstraction to classes may be helping with data sparseness, it is also throwing in a vast amount of noise, often so much that the relevant relationships are overwhelmed.

(175) a. Another major trick in making a portfolio recession-resistant is *choosing stocks in* "defensive" *industries*.
b. Big *investments in* "domestic" *industries* such as beer will make it even tougher for foreign competitors to crack the Japanese market.
c. The people with a *stake in* Nevada's gambling *industry* believe that they have barely tapped the potentially huge family trade.

Consider example (175a). Although the prepositional phrase *in industry* never occurs attached to either *choose* or *stock* in the WSJ training data, evidence from sentences like (175b) and (175c) provide evidence that it can be attached to other nouns like *investment* and *stake* having something in common with *stock* — something that is captured by the following scores:

v	c1	p c2	S^v	S^n
choose	⟨asset⟩	in ⟨enterprise⟩	19.85	839.44
choose	⟨asset⟩	in ⟨organization⟩	16.36	1084.75
choose	⟨asset⟩	in ⟨social_group⟩	15.02	1184.62

Unfortunately, the choice of attachment in this instance also ends up being influenced by a host of other completely irrelevant senses of *industry*. These include interpretations synonymous with the quality of industriousness, or with the activity of making goods rather than the organizations that engage in that activity.

c1	p c2	S^v	S^n
⟨accumulation⟩	in ⟨industriousness⟩	32.40	5.04
⟨asset⟩	in ⟨trait⟩	217.04	193.79
⟨change_of_magnitude⟩	in ⟨group_action⟩	550.55	534.99
⟨change_of_magnitude⟩	in ⟨attribute⟩	1087.57	1072.66
⟨increase⟩	in ⟨abstraction⟩	1302.13	1291.68

Although in this case conceptual association makes the correct choice despite such interference, irrelevant classes are having a significant impact on the experimental results.

A second undesirable effect of the paired t-test is the impact that the number of different class memberships for n2 has on confidence. The significance of a given value of t is calculated according to the number of degrees of freedom in the data — the larger the sample, the more degrees of freedom, and the lower t has to be in order to achieve significance. As the previous example illustrates, irrelevant class memberships already bring noise to the comparison between the two attachment sites; simply by virtue of their number they also tend to inflate confidence.

Finally, the use of mutual information as an association measure, and the weighting of the mutual information score in order to bias the computation in favor of large counts, warrant further consideration — mutual information has been criticized for, among other things, its poor behavior given low frequencies, and alternative measures of association may prove better.

On the positive side, it is clear that class information is providing some measure of resistance to sparseness of data. As mentioned earlier, adding the object of the preposition *without* using noun classes leads to hopelessly sparse data — yet the performance of the conceptual association strategy is far from hopeless. In addition, examination of what the conceptual association strategy actually did shows that the relationships being discovered are intuitively plausible — as for example in (175a), above, where *stock*, *stake*, and *investment* could all reasonably be viewed as assets or resources. Similarly, although *staff* belongs to 25 classes in WordNet — including ⟨musical_notation, 2332528⟩ and ⟨rod, 1613297⟩, for instance — a *staff in Dallas* is consistently interpreted as describing a group of personnel or people.

I would argue that the quantitative and qualitative facts, taken together, show that conceptual association is a good starting point for further work on broad-coverage application of class-based statistical disambiguation strategies. The central obstacle to improved performance appears to be ambiguity of class membership, and determination of class membership is a topic that shows signs of yielding to broad-coverage statistical techniques (Gale, Church, and Yarowsky, 1992a; Yarowsky, 1992; Yarowsky, 1993; Dagan and Itai, to appear).

5.4.5 Relation to other work

A number of other researchers have reached much the same conclusions presented here: there is a great deal of advantage to be gained from combining corpus-based lexical relationships with a model of word class membership. In the concluding part of this section I discuss several examples of such work being applied to problems in ambiguity resolution.

Weischedel *et al.* (1989; 1991) have investigated the use of a hand-constructed, domain-specific taxonomy together with corpus data from the MUC (Message Understanding Conference) evaluations for resolution of prepositional phrase attachment ambiguity. Their methodology has been to manually annotate nouns and verbs in a training sample with semantic tags from the taxonomy — for example, given "exploded at dawn," to annotate *dawn* with ⟨`time`⟩ and *explode* with ⟨`explosion event`⟩. Frequencies of syntactically-mediated lexical co-occurrences, expressed as relational triples, are then estimated either using parses in the Penn Treebank (Weischedel et al., 1989) or using partial parses produced by the MIT Fast Parser (Weischedel et al., 1991). Given lexical frequencies, semantic annotations, and a domain taxonomy, conditional probabilities are calculated using a "backing off" procedure: the probability of attaching a prepositional phrase P,O to attachment site X is calculated directly from the lexical frequencies, if any {X,P,O} co-occurrences are available in the training sample, and otherwise generalizations of X and O in the taxonomy are considered, with each generalization incurring a penalty. In addition, a probabilistic "closest attachment" heuristic is implemented by computing the probability p(d) that d words separate the head word X from the phrase to be attached. On a test set of prepositional attachments *not* made by the partial parser, Weischedel *et al.* report an accuracy of of 66% for the semantic model alone, 75% for the closest attachment model alone, and 82% for the two combined (by simply multiplying probabilities of the two models). No figures on the coverage-accuracy tradeoff are reported.

Although the work described by Weischedel *et al.* is very similar in spirit to the work described here, there are a number of important differences in the details. First, the conceptual model used is one that was designed specifically for the domain, and, though its size is not reported, is almost certain to be smaller and probably less fine-grained than the WordNet noun taxonomy. A second difference is the method by which words are mapped to classes in the taxonomy: manual annotation of each word with a unique semantic tag effectively solves the word sense disambiguation problem in advance. This may be important, though there is some reason to believe that within such a restrictive context the word sense problem would be significantly more constrained in any case (Gale, Church, and Yarowsky, 1992b). A third difference is the combination in their work of a lexical or conceptual preference strategy with a purely structural strategy, a choice that appeared to have a significant effect on the results. This is something that should be considered in future work on conceptual association, since even (Whittemore, Ferrara, and Brunner, 1990) found right association to be useful as a fallback strategy when lexical preference was inconclusive.

Finally, and perhaps most interesting, is the question of finding appropriate levels of generalization within the taxonomy. Weischedel *et al.* (1989) comment that, given their manual annotation of concepts, "the critical issue is selecting the right level of generalization given the set of examples in the supervised training set" (p. 30). As discussed briefly in Chapter 4, Section 4.6.2, the use of mutual information in the context of an IS-A taxonomy has the interesting behavior of seeking a class that is general (increasing p($c|x$)) but not too general (or else p(c) will dominate). Thus using the measures proposed here, generalization to an appropriate level of abstraction may be happening as a side-effect rather than as the result of an explicit procedure designed for that purpose — notice, for example, the tendency to classify *stock* as ⟨`asset`⟩ given the evidence in example (175) rather than its subordinate ⟨`working_capital`⟩ or the superordinate

⟨possession⟩. It would be interesting to compare the results of such an "automatic" generalization process with the adaptation of Katz's (1987) backing off procedure that was used by Weischedel *et al.*

Grishman and Sterling (1992) report on the acquisition of "semantic patterns" using methods quite similar to Weischedel *et al.*: they employed a robust parser to extract relational triples, and used a set of manually prepared, corpus-specific word classes to generalize those triples under the assumption that each word is mapped to a unique, most specific class. A direct comparison with (Weischedel et al., 1991) and the work reported here is difficult, however, since Grishman and Sterling adopted different evaluation criteria. In one evaluation, they used relational triples extracted from training data as a filter on triples from test data; this permitted the computation of precision and recall for the filtered set as evaluated against human judgements on the same data. In another evaluation, they used the extracted relational triples to filter parser output, and compared the resulting parses against "correct" parses in the Penn Treebank. It is interesting to note one qualitative similarity between Grishman and Sterling's results and the results reported in this chapter: they report that generalizing to classes yields higher recall, but decreases accuracy.

In later work, Grishman and Sterling (1993) have shifted their method of generalization from hand-constructed semantic classes to a smoothing technique. (This was discussed in Chapter 2, Section 2.3.1.) Their comparison of the new (smoothing) method with the old (class-based) does not show any conclusive differences between the two.

(Chang, Luo, and Su, 1992) report on a model that combines semantically based co-occurrence probabilities with syntactic and lexical probabilities in a single unified framework. Unlike Grishman and Sterling, they do not localize relevant co-occurrences by extracting a set of relational triples; instead, they adopt an annotated context-free formalism in which semantic co-occurrences are made local by percolating semantic features up the tree.

(176) a. [VP(sta,anim) saw(sta) [NP(anim) the boy [PP(loc) in the park]]]
b. [VP(sta,loc) saw(sta) [NP(anim) the boy] [PP(loc) in the park]]

So, for example, the structure in (176a) would result in semantic features STA (presumably "stative") and ANIM ("animate") co-occurring at the VP node, whereas the VP in structure in (176b) would be annotated with a co-occurrence of stative and locative. In addition, the probabilistic model of information at a node is not strictly context-free: it takes into account a limited amount of surrounding context, as well. (This appears to be similar to the probabilistic model proposed in (Magerman, 1993).) Semantic features are taken from a set of 104 semantic tags (22 of them for nouns), and each word was apparently given a single unique tag as part of its lexical representation.

Chang *et al.* evaluated the contribution of semantic co-occurrences to ambiguity resolution by testing their model with and without the semantic score, with the evaluation criterion being the frequency with which the correct parse has the highest probability. Adding semantic scores to the syntactic model increased the percentage correct from 43% to 58%, about a 35% improvement. It is worth noting that this test was conducted on an extremely small sample, given the number of parameters in the model — they used 10-fold cross-validation on a sample of 1000 sentences.

(Basili, Pazienza, and Velardi, 1991) describe an investigation of prepositional phrase attachment in the context of a more general research program on the combination of natural language processing and statistical methods for lexical acquisition, an effort sharing many of the motivations of the present work (Velardi, 1991; Velardi, Pazienza, and Fasolo, 1991; Basili, Pazienza, and Velardi, 1992). Like Weischedel *et al.*, their method requires human intervention in order to augment a training corpus with annotations from a small, largely domain-dependent set of semantic tags, and, like Grishman and Sterling, a robust syntactic analyzer is used to extract tuples of lexical items co-occurring in a variety of syntactic relationships.

Judgements of prepositional phrase attachment are made using a measure that Basili *et al.* call "conditioned mutual information":

$$\mathrm{I}(X/prep, C) \quad = \quad \rho \frac{\mathrm{freq}(X, prep, C)}{\mathrm{freq}(prep, C)\mathrm{freq}(prep, X)}. \tag{5.10}$$

(A comparison of this score with the standard information-theoretic definition of conditional mutual information (Cover and Thomas, 1991, p. 22) makes it clear that $\rho = \mathrm{freq}(prep)$, which is constant for any given disambiguation decision.) Intuitively, the score measures the association of the attachment site X and the class C, given that they are related by *prep*. This differs somewhat from the intuition followed in Section 5.4.3, where the mutual information $\mathrm{I}(X; prep, C_2)$ measures the association between an attachment site and the prepositional phrase as a whole (not taking the preposition as given); more significantly, Basili *et al.* do not generalize from words to classes for nouns serving as the potential attachment site, only for the object of the preposition. Despite these differences in detail, the idea behind the association measure is quite similar to the one that was developed independently here.

Results are reported for experiments in two different text genres, one in the commercial domain and the other in the legal domain. Lexical association was adopted as a basis for comparison using a training sample extracted from the Italian text by their shallow parsing method.[12] The results indicate that their word-class association strategy is a useful one: in the commercial domain, an accuracy of 84% was achieved as compared to 74% for lexical association, and in the legal domain, an accuracy of 68% was achieved compared to only 49% for lexical association. There is no discussion on assessment of confidence or the coverage-accuracy tradeoff.

To summarize, the work reported in this chapter bears some important similarities to its predecessors, as well as a number of interesting differences. From a practical or methodological perspective, all the related work described in this section is predicated on the idea that class-based (semantic) relationships can provide a measure of robustness against sparse data, and, furthermore, that surface-structure parsing provides an adequate level of analysis for collecting data about semantic co-occurrences. In addition, from a linguistic standpoint, the authors discussed here seem to agree that the utility of lexical preference as a disambiguation strategy arises not strictly from lexical relationships, but from the underlying semantic relationships they encode. Finally, Basili *et al.* make use of an association measure quite similar to the one proposed here.

The investigation in this chapter differs from most related work in its commitment to broad-coverage knowledge sources and its avoidance of restricted domains. Although WordNet is an imperfect knowledge

[12] Basili *et al.* do not say how they resolved ambiguous cases in the training data. However, on p. 7 they comment that the t-score method "requires a domain-dependent morphologic lexicon augmented with syntactic expectations (complement structure of verbs)"; this suggests that they viewed Hindle and Rooth's disambiguation heuristics as part of the strategy and used them on the training data in their own experiments.

source, it does provide a great deal of useful word class information, and the techniques developed here make use of them without requiring manual encoding in the lexicon (Chang, Luo, and Su, 1992; Grishman and Sterling, 1992) or, worse, manual annotation of the training corpus (Basili, Pazienza, and Velardi, 1991; Weischedel et al., 1991). Furthermore, unlike most of the related methods, the mapping from words to classes need not be a unique one, in terms of either word senses or levels in the taxonomy. Instead, as noted earlier, the association score appears to accomplish something resembling sense disambiguation and appropriate generalization in the taxonomy as a side-effect (though this requires further study). Finally, I have attempted to provide a clear sense of how a class-based strategy affects the tradeoff between coverage and accuracy, rather than simply reporting the percentage of correct responses. The conceptual association strategy makes this possible by providing a measure of confidence along with its guess as to the correct choice.

5.5 Nominal Compounds

5.5.1 Syntactic bias and semantic preferences

Nominal compounds are another kind of "every way ambiguous" construction that appears repeatedly in most samples of unconstrained text. The phrase *water meter cover adjustment screw* has 14 possible bracketings, and it is not difficult to come up with enough additional modifiers (*Penn engineering building basement water meter cover adjustment screw*) to produce a not-unnatural phrase with an entirely staggering number of analyses (1430, in this case!). Furthermore, it is often difficult to make confident judgements about what analysis is correct for compound nominals — the Penn Treebank avoids the problem by not encoding NP-internal structure.

In this last section I will briefly develop a proposal from (Marcus, 1980) for resolving complex nominal compounds using a combined syntactic and semantic strategy. At the heart of Marcus's proposal are two hypotheses: first, that there is a syntactic bias in favor of immediately combining adjacent nouns, and second, that complex compounds can be handled iteratively looking at no more than three nouns at a time. The algorithm he proposes is quite simple; I reproduce it here in its entirety.

- Given a noun phrase consisting of two nouns n1 and n2:
  - If [n1 n2] is semantically acceptable, then build [n1 n2]
- Given three nouns n1, n2, and n3:
  - If either [n1 n2] or [n2 n3] is not acceptable,
    then build the alternative structure;
  - Otherwise, if [n2 n3] is semantically preferable to [n1 n2],
    then build [n2 n3];
  - Otherwise, build [n1 n2].

The intuition behind the algorithm is to combine nouns according to semantic relationships when possible, but to use the syntactic bias when the semantic preferences are inconclusive. Examples worked through by hand seemed to support this intuition, but Marcus was unable to go any further: at the time he proposed the algorithm, there was no way to measure the relevant semantic preferences. He wrote, "Because I know of

no technique which can answer the necessary semantic questions, this procedure has not been implemented" (p. 251).

5.5.2 Implementation

The selectional association between nominal modifiers and their heads, used in Section 5.3, provides a possible solution to the problem Marcus encountered. Equations (5.4) and (5.5), repeated here for convenience, are intended precisely to answer the semantic questions asked in Marcus's algorithm.

$$\mathrm{A}(n_m \rightarrow c_h) = \frac{\mathrm{p}(c_h|n_m) \log \frac{\mathrm{p}(c_h|n_m)}{\mathrm{p}(c_h)}}{\sum_c \mathrm{p}(c|n_m) \log \frac{\mathrm{p}(c|n_m)}{\mathrm{p}(c)}}. \tag{5.11}$$

$$\mathrm{A}(c_m \leftarrow n_h) = \frac{\mathrm{p}(c_m|n_h) \log \frac{\mathrm{p}(c_m|n_h)}{\mathrm{p}(c_m)}}{\sum_c \mathrm{p}(c|n_h) \log \frac{\mathrm{p}(c|n_h)}{\mathrm{p}(c)}}. \tag{5.12}$$

The following auxiliary definitions are also helpful:

$$\begin{aligned} \mathrm{A}(n1 \rightarrow n2) &= \max_{c_i} \mathrm{A}(n1 \rightarrow c_i) \\ \mathrm{A}(n1 \leftarrow n2) &= \max_{c_j} \mathrm{A}(c_j \leftarrow n2) \\ \mathrm{A}(n1, n2) &= \max \{\mathrm{A}(n1 \rightarrow n2), \mathrm{A}(n1 \leftarrow n2)\} \end{aligned} \tag{5.13}$$

That is, the selectional association between two nouns is based on the maximum word-to-class selectional association, taken in either direction over all possible classes. Given this definition, selectional association provides "subroutines" for Marcus's algorithm:

- A phrase [n m] is *semantically not acceptable* if $\mathrm{A}(n, m) \leq 0$, and *semantically acceptable* otherwise.
- A phrase [n m] is *semantically preferable* to [m k] if $\mathrm{A}(n, m) - \mathrm{A}(m, k) > \tau$, where τ is a parameter.

With these subroutines serving to answer the semantic questions, Marcus's algorithm is easy to implement and evaluate.

5.5.3 Quantitative Evaluation

As a preliminary attempt at evaluating Marcus's algorithm, I extracted a sample of 200 noun-noun-noun compounds from the *Wall Street Journal* corpus in the Penn Treebank, and assigned one of the two possible bracketings to each, using the sentence the compound appeared in and the previous and following sentences as context. I omitted five of the cases, either because I simply could not arrive at a judgement or because a three-noun compound was not the correct syntactic structure. The *a priori* bias in the test set was 64.1% in favor of combining the first two nouns.

Selectional association was estimated using the same sample of noun-noun co-occurrences used in Section 5.3; this sample is disjoint from the test set, since noun-noun compounds were taken from noun phrases containing exactly two nouns, and noun-noun-noun compounds were taken from noun phrases containing exactly three. The results were as follows:

	Coverage (%)	Accuracy (%)
$\tau = 0.0$	80.5	65.6
$\tau = 1.0$	80.5	66.2
$\tau = 2.0$	80.5	68.2
$\tau = 3.0$	80.5	72.6
$\tau = 4.0$	80.5	70.1

The coverage figure (157 of 195 examples) is constant because the algorithm in effect defaults to the syntactic strategy in the absence of other information; the only time there was no answer was when one of the nouns was unknown or when both bracketings were semantically unacceptable. Unknown words are by far the biggest culprit, with 18.5% of the examples containing a noun not covered by WordNet Version 1.2 — in general, the unknown word is either a hyphenated compound (recall that I have restricted my attention to single nouns even though WordNet does include compounds), a gerund, or a piece of specialized terminology.

(177) a. college-bowl type competitions
b. real-estate loan portfolios
(178) a. bank consulting firm
b. proprietary operating system
(179) a. female hormone diethylstilbestrol
b. retinoblastoma suppressor gene

In order to provide a baseline against which to evaluate these results, I had the test set (without contexts) bracketed by an independent judge, and in each case had the judge include a measure κ of confidence in the choice on a scale from 0 (not at all confident) to 4 (very confident). These compared to my judgements as follows:

	Coverage (%)	Agreement (%)
$\kappa \geq 0$	100.0	80.0
$\kappa \geq 1$	96.4	80.9
$\kappa \geq 2$	90.8	81.4
$\kappa \geq 3$	65.6	83.6
$\kappa \geq 4$	15.9	100.0

As in the case of prepositional phrase attachment, the quantitative results are equivocal. On the one hand, the best performance of the algorithm is only a 13% improvement over simply guessing the first bracketing. On the other hand, this is fully half the way to the 26% improvement realized by the human judge, if a generous attitude is taken about unknown words.

5.5.4 Qualitative Evaluation

A qualitative evaluation of the algorithm is quite informative. Table 5.4 shows the test examples that the algorithm mistakenly bracketed as [n1 n2] n3 when they should have been n1 [n2 n3]; Table 5.5 shows the converse, examples that were mistakenly bracketed as n1 [n2 n3] rather than [n1 n2] n3. (These are errors when $\tau = 3.0$.)

Example	A$(n1, n2)$	A$(n2, n3)$	Δ
bankers acceptance rate	38.02	2.53	-35.49
state law enforcers	5.43	0.07	-5.36
sports cable channel	4.70	3.98	-0.71
winter ski season	3.42	2.72	-0.70
exchange trading practices	2.97	2.41	-0.56
estate investment trust	3.44	3.08	-0.36
management information system	2.63	2.32	-0.30
investment trust funds	3.08	2.82	-0.25
executive dining room	0.15	0.04	-0.11
world business competition	2.86	2.76	-0.10
adult trade books	2.45	2.51	0.06
oil maintenance schedule	2.49	2.55	0.06
merchandise trade deficit	4.12	4.37	0.25
% sales tax	1.49	1.88	0.38
sample leave policies	2.27	2.95	0.67
college entrance examination	1.88	2.61	0.73
state housing prices	0.88	1.70	0.81
bank trade associations	1.69	2.51	0.81
takeover stock traders	1.58	2.44	0.85
farm price index	1.21	2.09	0.87
home fitness equipment	0.50	1.65	1.15
business trade groups	1.96	3.32	1.36
world oil markets	0.65	2.08	1.43
% sales boost	1.49	3.07	1.57
performance plastic materials	0.46	2.22	1.76
record trade deficit	2.46	4.37	1.90
state securities group	1.81	4.65	2.83
market interest rates	5.35	8.22	2.86

Table 5.4: Incorrect bracketings of [n1 [n2 n3]]

Example	A$(n1, n2)$	A$(n2, n3)$	Δ
bone marrow transplants	0.00	3.19	3.19
beauty product line	-0.06	3.16	3.23
shareholder rights plan	1.52	5.29	3.76
education services company	1.40	5.26	3.86
cash interest bill	1.59	5.61	4.01
motor vehicle maker	4.20	9.02	4.81
chief executive officer	11.96	16.80	4.83
semiconductor marketing arm	2.95	7.84	4.89
insurance brokerage agency	5.22	10.21	4.98
restaurant franchise system	2.31	7.37	5.06
farm income records	1.21	7.58	6.36
golf club makers	1.69	9.02	7.32
printer marketing arm	0.00	7.84	7.84
commodity brokerage firms	0.02	12.20	12.17
state loan guarantees	0.56	16.57	16.00

Table 5.5: Incorrect bracketings of [[n1 n2] n3]

The general pattern that emerges is one in which the general direction of the algorithm is correct, but subtleties are missed. For most of the cases in Table 5.4, there is in fact a semantic preference for [n2 n3], but the difference is not great enough to pass the threshold. For many of these, the [n1 n2] combination is entirely plausible — e.g. *oil maintenance, farm price, home fitness.* In Table 5.5, the pattern is strikingly different. In general the [n2 n3] combination overwhelms [n1 n2], and in many cases this appears to be justified — for example, *loan guarantees, brokerage firms, income records, marrow transplant,* and *product line* are all tight collocations. The lesser association of the [n1 n2] combinations in these cases — *state loan, commodity brokerage, farm income, bone marrow, beauty product* — tends to arise because the collocations have low frequency in the corpus, even when class membership is taken into account. In some cases, inappropriate classes are coming into play: *income* and *record* associate so strongly because ⟨`record`⟩ IS-A ⟨`document`⟩ IS-A ⟨`possession`⟩, and collocations like *income return, income security,* and *income tax* lead to a very high value for A$(income \rightarrow$ ⟨`possession`⟩$)$.

There is a good chance that many of these problems will disappear when larger corpora are used and when word senses are taken into account. However, inspection of the incorrect choices also exposes some deeper problems. Consider the phrase *winter ski season*. The training data contain a number of instances for which a nominal modifier of *season* is a sporting activity — *baseball season, hunting season, skiing season* — and the association measure captures this generalization, since in each of these cases the highest scoring class for the modifier is ⟨`sport`⟩. However, interpreting *ski* as a modifier in this category requires inferring the relevant relationship between *ski* and *skiing*, something that may be possible in this case (perhaps via the relationship between a sport and its equipment?) but would probably not be straightforward to do in general.

As a second example, consider the set of nominal modifiers for the head *product*. The modifiers of *product* in the training set fall, with few exceptions, into the following rough groupings:

- **X product** ⇒ product made from or consisting of X:
  basket carbon chemical cocoa cosmetic dairy drug egg film food hardware insulin life-insurance

mainframe oil paper petroleum plastic polystyrene semiconductor sheet software steel storage-case system tape textile tissue tobacco underwear

- **X product** ⇒ product used in activity X:
  biotechnology building business communication control dialysis health-care home-improvement information-processing investment packaging plant-science skin-care storage telecommunication

- **X product** ⇒ product used to produce condition X:
  animal-health fitness

- **X product** ⇒ product used in location X:
  farm home household hospital office

A difficulty with using selectional association to assess semantic preference is that there may not be a direct mapping between groupings of this kind and classes in the taxonomy. Using selectional association, the first group of modifiers tends to be categorized as ⟨`object`⟩, which seems to be a reasonable fit, but no single class predominates in this way for the other groups. For example, in the second group words that can be interpreted both as activities and as objects (e.g. *building*, *business*, *control*) get lumped in with ⟨`object`⟩, and the remainder scattered among relatively weakly scoring classes like ⟨`communication`⟩ and ⟨`activity`⟩.

The problem is reflected in the algorithm's error on *beauty product line*. Although under a suitable interpretation, *beauty* could be interpreted as a member of the third group — a condition caused by the product — no such interpretation is available on the basis of the classes to which *beauty* belongs. Indeed, the error would remain even if beauty were a member of class ⟨`condition`⟩: even though *health* and *fitness* are both members of that class, the selectional association A(⟨`condition`⟩ ← *product*) is quite low.

What appears to be needed here, then, is an understanding of the semantic connection *underlying* the modifier-head relationship. This is something that it is difficult to imagine ascertaining automatically (though see (Basili, Pazienza, and Velardi, 1991; Velardi, 1991) for some discussion of partially automating the process). To echo the quote from (Hindle and Rooth, 1993) at the beginning of Section 5.4, if deeper semantic relationships of this kind are necessary in general, then it is hard to see how computational models are going to be able to solve this problem in unrestricted text any time soon.

Fortunately, this may not be the case. In the current example, although *beauty product line* is bracketed incorrectly, three other items in the test set are bracketed correctly — *cement products company*, *food products concern*, and *forest products company*. Although further experimentation is necessary, these examples encourage me to believe that the syntactic head-modifier relationship, mediated by conceptual classes, will suffice more often than not.

Chapter 6

Conclusions

6.1 Contributions

The core of the dissertation, in Chapter 3, is a new formalization of selectional constraints in information-theoretic terms. I think the proposal to treat selectional constraints from an inferential point of view, but to "hide" inference within the semantics of a taxonomic representation, is a novel one. In addition, I think that the time is right to have revived the question of how "information," in the sense of Shannon and Weaver (1949), is related to semantic content, as discussed by Bar-Hillel (1964), and to the process of interpretation.

The main contribution of Chapter 4 is a new account of one particular kind of diathesis alternation, an account that, unlike most discussions of verbal diathesis, focuses on the verb-argument relationship rather than on a particular semantic property of the verb. The computational experiments in that chapter suggest that the model of selectional preference proposed here captures important aspects of inferability for argument properties; in that sense I think it lays the groundwork for a new set of mathematical and computational proposals about on-line processing, consistent with discussions of processing in terms of probabilistic constraints. (See also the discussion of argument plausibility in Chapter 3.) I think this model may also shed new light on the process of verb acquisition, since bootstrapping proposals are increasingly coming to recognize the importance of argument properties in that process.

In Chapter 5, I think I have demonstrated the utility of using knowledge-based classes in syntactic disambiguation by statistical methods. Furthermore, unlike most statistical approaches, disambiguation relies not on just any statistical test, but on an association measure that was independently motivated and justified in the previous chapters. Admittedly this may not be of great concern in practical applications, but at the very least I have provided a starting point, and a set of initial results, for statistical approaches that make use of a broad-coverage taxonomy in unconstrained text.

Finally, the underlying premise of this work has been that the information-theoretic view of language as a stochastic phenomenon and the linguistic view of language as a cognitive phenomenon, though often characterized as being in opposition to each other, are not fundamentally incompatible. I believe that the results in the preceding chapters support this conclusion, and I hope the thesis as a whole will contribute to making it more widely accepted.

6.2 Thoughts on Future Work

Speculating about possible future directions is extremely easy, since I have done my best to relate this work to many different areas of intellectual pursuit. In this section, therefore, I will just briefly mention three directions that strike me as particularly interesting.

Word sense disambiguation. One issue that came up repeatedly throughout the thesis was word sense disambiguation — both with regard to the training data, and with regard to the use of selectional constraints as a way of identifying the most plausible reading of a word in its context as the argument of a predicate. One straightforward thing to do would be to abandon the uniform distribution of credit among noun classes — e.g., observing *drink wine* and incrementing ⟨color⟩ and ⟨beverage⟩ by equal amounts — and instead to use an existing word-sense disambiguation technique to obtain a better approximation of how credit should be distributed. An alternative, suggested to me by David Magerman, would be to construct a hidden model in which observed predicate-word co-occurrences provide the data for re-estimation of predicate-class probabilities using the EM algorithm.

The relationship between selectional constraints and lexical disambiguation has been in evidence at least since (Katz and Fodor, 1964), and the behavior of the implemented model suggests that selectional association in many cases provides strong evidence for a particular sense. This finding is consistent with Yarowsky's (1993) claim that local collocational relationships provide a reliable source of evidence for sense disambiguation when such relationships are present. Yarowsky suggests that class-based collocations may help resolve some of the problems his method encounters with low recall, and the integration of the present approach with his technique seems well worth pursuing.

Basic levels. A difference between typical word-sense disambiguation methods and selectional association is that, where senses are typically selected from a "flat" set, the classes under consideration in the present work are part of a multi-level taxonomy. It has been widely noted that the selection of an appropriate level of abstraction is a difficult problem — for example, Velardi *et al.* (1991, p. 164) comment, "The most difficult task . . . is to define at the appropriate level of generality the selectional restrictions on conceptual relations."

A small but significant contribution of the thesis is that the measure of selectional association locates an "appropriate" level within the taxonomy automatically, by trading off a marginal class probability (which goes up as you go higher in the taxonomy) against a conditional class probability (which decreases if you go too high). In future work I would like to investigate this property of the measure further, and to explore the possibility that it is related to the notion of basic level categories.

Underlying semantics. Although in Section 2.4.1 I attempted to provide a reasonable discussion of the semantics behind the taxonomy, and particularly its relationship to inference, further work on this topic is needed. Steve Abney points out some necessary elaborations to the notions of "plausible entailment" and "representative sentence" — at a minimum, the discussion should be expressed in terms of open propositions rather than open sentence frames (e.g. $\lambda f[\exists x(f(x)\&\text{sawed-in-two}(j,x))]$ rather than "John sawed a ___ in two"); the notion of "representative sentence" needs to be formulated so as to exclude such cases as $\lambda f[\text{member-of}(f,\{f_1,f_2\})]$ (otherwise my criteria would let any pair of senses be synonyms); and there needs to be a clearer characterization of what would be excluded as a "plausible entailment" of a proposition (to prevent the criterion for synonymy from being too strict). The challenge in this task is to make the definitions more formal while at the same time not requiring a complete formalization of human inference.

Appendix A

Notes on Probability Estimation

A.1 Unit Credit Assignment

Although equation (2.14) represents the correct formalization of joint class-based probabilities, in earlier versions of this work (Resnik, 1992a; Resnik, 1992c; Resnik, 1993; Resnik and Hearst, 1993) and in Chapter 5 I used the following frequency estimate:

$$\text{freq}(x, c) \quad = \quad \sum_{w \in words(c)} \text{freq}(x, w). \tag{A.1}$$

That is, the joint frequency with x was increased by a unit rather than a fractional amount for each class to which w belonged. This is technically incorrect as far as the probabilistic model is concerned, since it leads to $\hat{p}_{MLE}(v, c)$ not being a probability function — notice how the marginal probability of x will be inflated for those x that tend to appear with nouns belonging to many different classes.

In the next section, I describe in detail how probability estimation was actually carried out in the work using that frequency estimate, in particular the use of the Good-Turing estimate rather than MLE. In the section that follows, I work through a simple example to illustrate how frequency and probability estimates are done now.

A.2 Good-Turing Estimates

Together with the frequency estimate in (A.1), in earlier experiments I used the Good-Turing (GT) estimator of probabilities (Good, 1953). The GT estimate is calculated by organizing observations in the sample according to frequency, so that bin n_r represents the number of items that were observed exactly r times; for example, n_2 is the number of items that were observed exactly twice. If k is the maximum number of times any item was observed, then

$$\sum_{r=1}^{k} r n_r = N, \tag{A.2}$$

where N is the total size of the sample. In order to estimate the probability of something that occurred r times, the maximum likelihood estimator would simply use the normalized frequency, i.e.

$$\hat{p}_{MLE} \quad = \quad \frac{r}{N}, \tag{A.3}$$

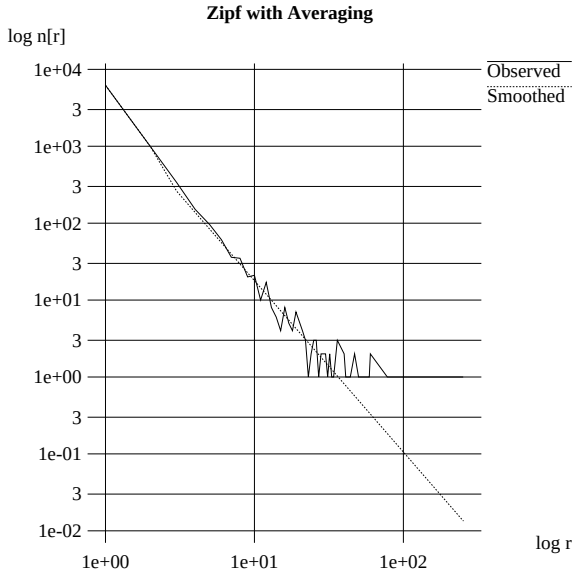


Figure A.1: Example of smoothing in Good-Turing probability estimation

but in contrast, the Good-Turing estimate is calculated by first computing an *adjusted* frequency r^*:

$$r^* = (r+1)\frac{n_{r+1}}{n_r}. \tag{A.4}$$

It is this adjusted frequency that is then normalized in order to estimate the probability. That is, the estimated probability for something that occurred r times in the sample is given by

$$N^* = \sum_{r=0}^{k} r^* n_r \tag{A.5}$$

$$\hat{p}_{GT} = \frac{r^*}{N^*}. \tag{A.6}$$

In practice, it is necessary to smooth the n_r — notice that if by chance no item in the sample had an observed frequency of exactly r, the denominator in equation (A.4) would be zero. Having observed that a plot of $\log n_r$ versus $\log r$ was very nearly linear, I experimented with smoothing the n_r by fitting the observed data to the equation

$$\log n_r = -m(\log r) + b. \tag{A.7}$$

As it turns out, this is equivalent to saying that

$$n_r \propto \frac{1}{r^m}, \tag{A.8}$$

which is to say that the class distribution follows Zipf's law.[1] Figure A.1 shows an example of how frequency estimates were smoothed.

In the interest of fidelity to a probabilistic framework, I have redone most of the experiments in this thesis using the frequency estimate in equation (2.14); this permits a true information-theoretic interpretation of the proposal made in Chapter 3. (Maximum likelihood is used unless otherwise noted.) In the next section I

[1] I am grateful to Ken Church for pointing this out; for further smoothing subtleties see (Church and Gale, 19xx, pp. 8-9).

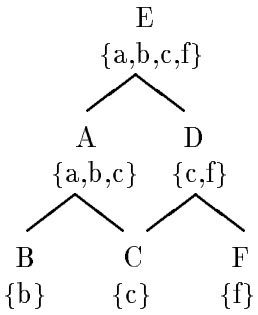


Figure A.2: A simple taxonomy

work through a small example to illustrate exactly how the frequency estimation is carried out and how the probability estimates are related to the structure of the taxonomy.

In future work, I may shift from MLE to the Cat-Cal estimator proposed in (Church and Gale, 19xx), since it requires fewer statistical assumptions than Good-Turing and is better equipped to dealing with fractional counts. However, I should note that changes in probability estimates do not appear to lead to any great differences in the reported results. I have found that experiments produce comparable results regardless of the probability estimator used.

A.3 Frequency Estimates Using the Taxonomy

As mentioned in Section 2.4.2, the structure of the taxonomy plays a role not in the formalization of the sample space, but in the estimation of the probability function. In order to examine this in a bit more detail, I will work through a simple example.

Consider the taxonomy in Figure A.2. Capital letters represent the set of class labels, {A,B,C,D,E,F}, and the sets below each label represent the extension of each class, words represented in lowercase. Notice that each link in the taxonomy corresponds to a subset-superset relationship between those extensions.

Suppose that what is observed is a 4-word sample: *a, b, c, f*. Since word *a* belongs to two classes, A and E, each of those two classes will have its frequency incremented by $\frac{1}{2}$. Similarly, when *b* is observed, classes B, A, and E will each be incremented by $\frac{1}{3}$. The entire sequence leads to the following summary of frequency assignment:

	a	b	c	f
A	$\frac{1}{2}$	$\frac{1}{3}$	$\frac{1}{4}$	
B		$\frac{1}{3}$		
C			$\frac{1}{4}$	
D			$\frac{1}{4}$	$\frac{1}{3}$
E	$\frac{1}{2}$	$\frac{1}{3}$	$\frac{1}{4}$	$\frac{1}{3}$
D				$\frac{1}{3}$

This results in the following frequency and probability estimates:

Frequency	MLE
f(A) = $\frac{13}{12}$	p(A) = .2708
f(B) = $\frac{1}{3}$	p(B) = .0833
f(C) = $\frac{1}{4}$	p(C) = .0625
f(D) = $\frac{7}{12}$	p(D) = .1458
f(E) = $\frac{17}{12}$	p(E) = .3542
f(F) = $\frac{1}{3}$	p(F) = .0833

As expected, since there are six classes in the taxonomy, the probability space can be viewed as describing a six-sided die — on any roll of the die, there is for example a 0.1458 probability of coming up with D. Unlike a die, however, the possible outcomes are in fact related: as you move up in the taxonomy from subsets to supersets, the probability necessarily increases.

Admittedly, there is something counterintuitive about assigning class C a different probability than classes B and F, given that *b*, *c*, and *f* were each observed once. The uniform distribution of credit among classes for word observations is at best a brute-force method; in general, the question of how probability should be assigned to classes in a taxonomy warrants further attention than it has been given here.

Appendix B

Experimental Data from Chapter 4

B.1 Experiment 1, Brown Corpus

Object-drop verbs		Non-object-drop verbs	
Verb	Strength	Verb	Strength
pour	4.80	hang	3.35
drink	4.38	wear	3.13
pack	4.12	open	2.93
sing	3.58	say	2.82
steal	3.52	like	2.59
eat	3.51	hit	2.49
push	2.87	catch	2.47
pull	2.77	do	1.84
write	2.54	want	1.52
play	2.51	show	1.39
explain	2.39	bring	1.33
read	2.35	put	1.24
watch	1.97	see	1.06
hear	1.70	find	0.96
call	1.52	take	0.93
		get	0.82
		give	0.79
		make	0.72
		have	0.43

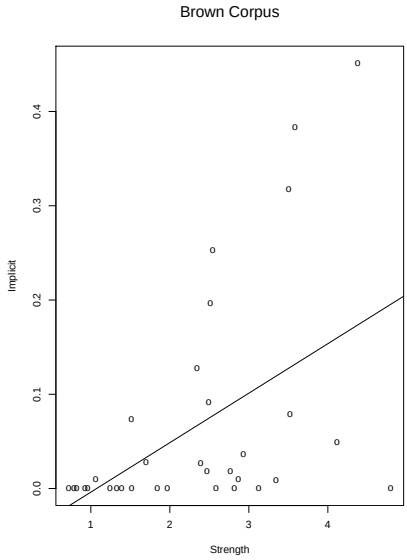
B.2 Experiment 1, CHILDES

Object-drop verbs		Non-object-drop verbs	
Verb	Strength	Verb	Strength
explain	4.41	open	2.41
pack	3.71	hang	2.03
sing	3.15	wear	2.02
read	2.58	show	1.83
drink	2.38	catch	1.67
write	2.33	hit	1.31
pour	2.30	give	1.18
steal	2.28	say	0.94
play	2.13	like	0.89
push	1.77	bring	0.88
hear	1.67	make	0.77
pull	1.55	take	0.74
watch	1.44	find	0.71
eat	1.15	want	0.70
call	0.95	see	0.48
		put	0.40
		get	0.28

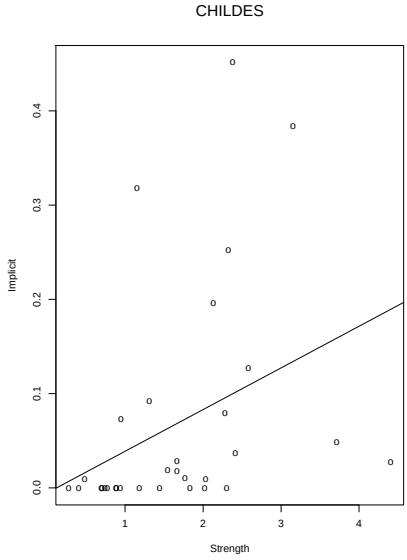
B.3 Experiment 1, Norms

Object-drop verbs		Non-object-drop verbs	
Verb	Strength	Verb	Strength
drink	2.83	say	2.56
play	2.64	wear	2.30
sing	2.63	do	2.21
pour	2.57	hang	1.96
eat	2.47	catch	1.92
call	2.39	hit	1.91
pull	2.22	open	1.88
explain	2.20	give	1.81
write	2.18	want	1.71
push	1.98	make	1.58
watch	1.86	see	1.54
read	1.81	show	1.42
pack	1.75	put	1.34
hear	1.71	like	1.30
steal	1.34	find	1.30
		take	1.28
		have	1.23
		get	1.17
		bring	1.04

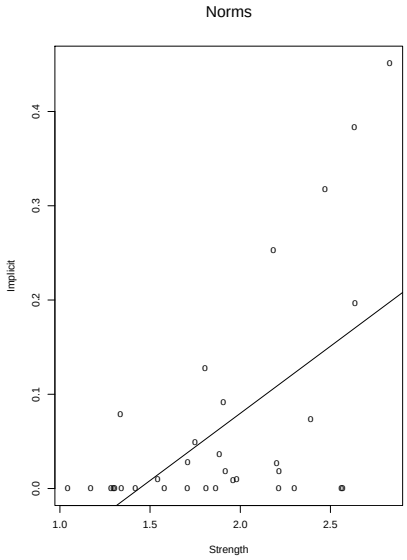
B.4 Experiment 2, Brown Corpus



B.5 Experiment 2, CHILDES Corpus



B.6 Experiment 2, Norms



B.7 Experiment 3, Verbs from Lehrer's (1970) Verb Classification

Type I: babysit, bale, breathe, conceive, cook, dance, date, draw, dream, drink, drive, eat, fly, hear, hum, iron, kick, marry, nod, paint, plow, print, read, reap, shrug, sing, smell, sow, spell, spit, swallow, think, type, wave, weave, write, yell

Type III: answer, approach, approve, attend, bid, build, call, change, choose, continue, discontinue, copy, cut, endure, enter, fail, follow, gain, govern, grab, guess, hoard, judge, know, lead, leave, lose, obey, disobey, order, pack, pass, pay, play, pour, promise, recall, refuse, remember, resist, spill, wash, waste, watch

B.8 Experiment 3, Brown Corpus

Type I verbs		Type III verbs	
Verb	Strength	Verb	Strength
dance	9.36	guess	5.89
spit	8.50	endure	5.41
yell	7.44	disobey	5.09
shrug	6.80	fail	5.02
reap	5.93	discontinue	4.98
sow	5.86	bid	4.86
date	5.86	pour	4.80
weave	5.48	refuse	4.40
hum	5.46	obey	4.20
type	5.09	pack	4.12
spell	5.04	waste	3.89
breathe	4.99	answer	3.60
plow	4.87	wash	3.47
iron	4.71	resist	3.35
cook	4.48	grab	3.32
drink	4.38	recall	3.31
swallow	4.07	judge	3.30
fly	3.99	attend	3.22
nod	3.76	promise	3.12
conceive	3.72	approve	3.08
kick	3.70	pay	2.85
smell	3.65	cut	2.79
wave	3.64	govern	2.67
sing	3.58	lead	2.56
eat	3.51	continue	2.52
marry	3.44	play	2.51
print	3.25	build	2.49
drive	3.12	remember	2.37
paint	2.94	order	2.35
think	2.56	choose	2.19
write	2.54	gain	2.17
read	2.35	approach	2.15
draw	1.95	change	2.05
hear	1.70	pass	2.01
		watch	1.97
		enter	1.81
		know	1.61
		follow	1.54
		call	1.52
		leave	1.48
		lose	1.47
		spill	0.00

B.9 Experiment 3, CHILDES

Type I verbs		Type III verbs	
Verb	Strength	Verb	Strength
dream	8.64	discontinue	6.97
shrug	6.85	continue	5.87
plow	5.24	gain	5.37
iron	5.09	approach	5.32
type	4.92	lead	4.42
spit	4.64	answer	3.90
nod	4.39	pack	3.71
wave	3.84	pay	3.58
fly	3.81	judge	3.35
smell	3.27	copy	3.05
drive	3.23	order	3.02
yell	3.20	choose	2.99
sing	3.15	waste	2.79
marry	2.85	pass	2.68
read	2.58	build	2.67
kick	2.40	promise	2.56
dance	2.40	guess	2.55
drink	2.38	follow	2.50
write	2.33	spill	2.37
spell	2.32	change	2.35
swallow	2.27	pour	2.30
paint	2.22	play	2.13
cook	2.01	grab	1.90
hear	1.67	wash	1.80
draw	1.60	cut	1.50
think	1.26	watch	1.44
eat	1.15	lose	1.33
		remember	1.23
		know	1.17
		call	0.95
		leave	0.81

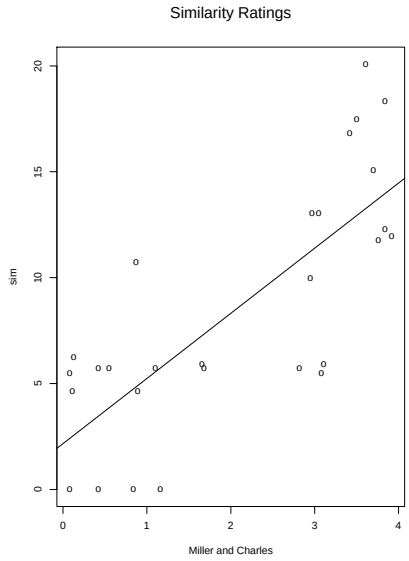
Appendix C

Word Similarity Data from Chapter 5

The following table gives the data from Miller and Charles's (1991) study, followed by their subjects' mean rating, the mean rating in my replication, and the similarity value calculated using equation (5.2).

Word Pair		Miller and Charles	Replication	sim
car	automobile	3.92	3.9	11.98
gem	jewel	3.84	3.5	18.34
journey	voyage	3.84	3.5	12.27
boy	lad	3.76	3.5	11.79
coast	shore	3.70	3.5	15.09
asylum	madhouse	3.61	3.6	20.08
magician	wizard	3.50	3.5	17.49
midday	noon	3.42	3.6	16.80
furnace	stove	3.11	2.6	5.90
food	fruit	3.08	2.1	5.47
bird	cock	3.05	2.2	13.06
bird	crane	2.97	2.1	13.06
tool	implement	2.95	3.4	9.96
brother	monk	2.82	2.4	5.74
crane	implement	1.68	0.3	5.74
lad	brother	1.66	1.2	5.90
journey	car	1.16	0.7	0.00
monk	oracle	1.10	0.8	5.74
cemetery	woodland	0.95	0.6	n/a
food	rooster	0.89	1.1	4.65
coast	hill	0.87	0.7	10.72
forest	graveyard	0.84	0.6	0.00
shore	woodland	0.63	0.7	n/a
monk	slave	0.55	0.7	5.74
coast	forest	0.42	0.6	0.00
lad	wizard	0.42	0.7	5.74
chord	smile	0.13	0.1	6.24
glass	magician	0.11	0.1	4.65
noon	string	0.08	0.0	0.00
rooster	voyage	0.08	0.0	5.49

The following plot illustrates the relationship between the Miller and Charles means and the calculated similarity value:



Bibliography

AHD. 1991. *American Heritage Dictionary*. Houghton Mifflin.

Adrian Akmajian and Frank Heny. 1975. *An introduction to the principles of transformational syntax*. MIT Press.

Robert B. Allen. 1990. Connectionist language users. *Connection Science*, 2(4):279–311.

Hiyan Alshawi. 1987. Processing dictionary definitions with phrasal pattern hierarchies. *Computational Linguistics*, 13(3-4):195–202.

Gerry Altmann and Mark Steedman. 1988. Interaction with context during human sentence processing. *Cognition*, 30:191–238.

S. Armstrong, L. Gleitman, and H. Gleitman. 1983. What some concepts might not be. *Cognition*, 13:263–308.

D. Ayuso, G. Donlon, D. MacLaughlin, L. Ramshaw, P. Resnik, V. Shaked, and R. Weischedel. 1989. A guide to IRUS-II application development. BBN Report 7144, Bolt, Beranek and Newman.

Lalit R. Bahl, Peter F. Brown, Peter V. de Souza, and Robert L. Mercer. 1989. A tree-based statistical language model for natural language speech recognition. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 37:1001–1008, July.

L.R. Bahl, F. Jelinek, and R.L. Mercer. 1983. A maximum likelihood approach to continuous speech recognition. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, PAMI-5:179–190.

Yehoshuah Bar-Hillel. 1964. *Language and Information*. Addison-Wesley.

Roberto Basili, Maria Teresa Pazienza, and Paola Velardi. 1991. Combining NLP and statistical techniques for lexical acquisition. In *Proceedings of the AAAI Fall Symposium on Probabilistic Approaches to Natural Language*, Cambridge, Massachusetts, October.

Roberto Basili, Teresa Pazienza, and Paola Velardi. 1992. Computational lexicons: the neat examples and the odd exemplars. In *Third Conference on Applied Natural Language Processing*, pages 96–103. Association for Computational Linguistics, March.

Richard Beckwith, Christiane Fellbaum, Derek Gross, and George Miller. 1991. WordNet: A lexical database organized on psycholinguistic principles. In Uri Zernik, editor, *Lexical Acquisition: Exploiting On-Line Resources to Build a Lexicon*, pages 211–232. Erlbaum.

Peter A. Bensch and Walter J. Savitch. 1992. An occurrence-based model of word categorization. Presented at 3rd Meeting on Mathematics of Language (MOL3), November.

L. Boggess, R. Agarwal, and R. Davis. 1991. Disambiguation of prepositional phrases in automatically labeled technical texts. In *Proceedings of AAAI-91*.

Julie Boland, Michael Tanenhaus, Greg Carlson, and Susan Garnsey. 1989. Lexical projection and the interaction of syntax and semantics in parsing. *Journal of Psycholinguistic Research*, 18(6):563–575.

Lisa Braden-Harder and Wlodek Zadrozny. 1991. Lexicons for broad coverage semantics. In Uri Zernik, editor, *Lexical Acquisition: Exploiting On-Line Resources to Build a Lexicon*, pages 369–388. Erlbaum.

L. Breiman, J. Friedman, R. Olshen, and C. Stone. 1984. *Classification and Regression Trees*. Wadsworth and Brooks.

Eric Brill, D. Magerman, M. Marcus, and B. Santorini. 1990. Deducing linguistic structure from the statistics of large corpora. In *DARPA Speech and Natural Language Workshop*, pages 275–282.

Eric Brill. 1991. Discovering the lexical features of a language. In *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics, Berkeley, CA*.

Paula Brown and Gary Dell. 1987. Adapting production to comprehension: the explicit mention of instruments. *Cognitive Psychology*, 19:441–472.

Peter F. Brown, Vincent J. Della Pietra, Peter V. deSouza, and Robert L. Mercer. 1990. Class-based n-gram models of natural language. In *Proceedings of the IBM Natural Language ITL*, pages 283–298, Paris, France, March.

Peter F. Brown, Vincent J. Della Pietra, Peter V. deSouza, Jennifer C. Lai, and Robert L. Mercer. 1992. Class-based n-gram models of natural language. *Computational Linguistics*, 18(4):467–480, December.

Wayles Browne. 1971. Verbs and Unspecified NP Deletion. *Linguistic Inquiry*, 2:259–260.

M. G. Bulmer. 1967. *Principles of Statistics*. Dover Publications.

Roy Byrd, Nicoletta Calzolari, Martin Chodorow, Judith Klavans, and Mary Neff. 1987. Tools and methods for computational linguistics. *Computational Linguistics*, 13(3-4):219–240.

Greg Carlson and Michael Tanenhaus. 1988. Thematic roles and language comprehension. In W. Wilkins, editor, *Thematic Relations*, volume 21 of *Syntax and Semantics*, pages 263–288. Academic Press.

Jing-Shin Chang, Yih-Fen Luo, and Keh-Yih Su. 1992. GPSM: A generalized probabilistic semantic model for ambiguity resolution. In *Proceedings of the 30th Annual Meeting of the Association for Computational Linguistics*, pages 177–184. ACL, June.

Noam Chomsky. 1957. *Syntactic Structures*. Mouton, The Hague.

Noam Chomsky. 1965. *Aspects of the Theory of Syntax*. MIT Press, Cambridge, MA.

Kenneth Church and William Gale. 19xx. Enhanced Good-Turing and Cat-Cal: Two new methods for estimating probabilities of English bigrams. ms.

K. Church and P. Hanks. 1989. Word association norms, mutual information, and lexicography. In *Proceedings of the 27th Meeting of the Association for Computational Linguistics*. Vancouver, B.C.

Kenneth Church and Robert Mercer. 1993. Introduction to the special issue on computational linguistics using large corpora. *Computational Linguistics*, 19(1):1–24.

Kenneth W. Church and Ramesh Patil. 1982. Coping with syntactic ambiguity or how to put the block in the box on the table. *American Journal of Computational Linguistics*, 8(3-4):139–149.

K. Church, W. Gale, P. Hanks., and D.M. Hindle. 1990. Using statistics in lexical analysis. In Uri Zernik, editor, *Lexical Acquisition: Using On-line Resources to Build a Lexicon*. Lawrence Erlbaum.

Kenneth Church, William Gale, Patrick Hanks, and Donald Hindle. 1991. Using statistics in lexical analysis. In Uri Zernik, editor, *Lexical Acquisition: Exploiting On-Line Resources to Build a Lexicon*, pages 116–164. Erlbaum.

Sharon Cote. 1992. Discourse functions of two types of null objects in English. Presented at the 66th Annual Meeting of the Linguistic Society of America, Philadelphia, PA, January.

Thomas M. Cover and Joy A. Thomas. 1991. *Elements of Information Theory*. John Wiley.

S. Crain and M. Steedman. 1985. On not being led up the garden path: the use of context by the psychological syntax processor. In D. Dowty, L. Karttunen, and A. Zwicky, editors, *Natural Language Processing: Psychological, Computational, and Theoretical Perspectives*. Cambridge University Press.

D. A. Cruse. 1986. *Lexical Semantics*. Cambridge University Press.

Ido Dagan and Alon Itai. (to appear). Word sense disambiguation using a second language monolingual corpus. To appear in Computational Linguistics.

I. Dagan. 1990. A statistical filter for resolving pronoun references. In *Proceedings of the 7th Israeli Symposium on Artificial Intelligence and Computer Vision*.

K. Dahlgren and J. McDowell. 1986. Using commonsense knowledge to disambiguate prepositional phrase modifiers. In *AAAI-86*, pages 589–593.

Scott Deerwester, Susan Dumais, Goerge Furnas, Thomas Landauer, and Richard Harshman. 1990. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6):391–407.

D. Dowty, R. Wall, and S. Peters. 1981. *Introduction to Montague Semantics*. D. Reidel Publishing Co., Boston.

David Dowty. 1991. Thematic proto-roles and argument selection. *Language*, 67(3):547–615.

Theodore Drange. 1966. *Type Crossings: Sentential Meaningless in the Border Area of Linguistics and Philosophy*. Mouton.

Jeffrey Elman. 1989. Representation and structure in connectionist models. Technical Report CRL TR-8903, University of California, San Diego.

Jeffrey Elman. 1990. Finding structure in time. *Cognitive Science*, 14:179–211.

U. Essen and V. Steinbiss. 1992. Cooccurrence smoothing for stochastic language modeling. In *ICASSP-92*, pages I–161 – I–164, San Francisco.

Christiane Fellbaum and Judy Kegl. 1989. Taxonomic structures and cross-category linking in the lexicon. In *Proceedings of ESCOL '89*, pages 93–104.

Evelyn Ferstl. 1993. The role of lexical information and discourse context in syntactic processing: a review of psycholinguistic studies. Technical Report 93-03, University of Colorado at Boulder.

Charles Fillmore. 1986. Pragmatically controlled zero anaphora. In *Proceedings of the Berkeley Linguistics Society*, pages 95–107.

Cynthia Fisher, Geoffrey Hall, Susan Rakowitz, and Lila Gleitman. 1994. When it is better to receive than to give: syntactic and conceptual constraints on vocabulary growth. *Lingua*, 92(1).

Cynthia Fisher, Lila Gleitman, and Henry Gleitman. 1991. On the semantic content of subcategorization frames. *Cognitive Psychology*, 23(3):331–392, July.

J. A. Fodor, M. F. Garrett, E. T. Walker, and C. Parkes. 1980. Against definitions. *Cognition*, 8(3):1–105.

Janet Dean Fodor. 1977. *Semantics: theories of meaning in generative grammar*. Harvard University Press.

Marilyn Ford, Joan Bresnan, and Ronald Kaplan. 1982. A competence-based theory of syntactic closure. In Joan Bresnan, editor, *The Mental Representation of Grammatical Relations*. MIT Press.

W. Francis and H. Kučera. 1982. *Frequency Analysis of English Usage*. Houghton Mifflin Co.: New York.

L. Frazier. 1979. *On Comprehending Sentences: Syntactic Parsing Strategies*. Ph.D. thesis, University of Massachusetts.

William Gale, Kenneth Church, and David Yarowsky. 1992a. A method for disambiguating word senses in a large corpus. Statistical Research Reports 104, AT&T Bell Laboratories, March. (To appear in *Computers and Humanities*).

William Gale, Kenneth Church, and David Yarowsky. 1992b. One sense per discourse. Proceedings of the 4th DARPA Speech and Natural Language Workshop, February.

G. Gazdar, E. Klein, G. Pullum, and I. Sag. 1985. *Generalized Phrase Structure Grammar*. Harvard University Press.

Jane Gillette and Lila Gleitman. forthcoming. Effects of situational cues on the identification of nouns and verbs. ms.

Lila Gleitman, Henry Gleitman, Barbara Landau, and Eric Wanner. 1988. Where learning begins: initial representations for language learning. In *Linguistics: The Cambridge Survey, Volume III: Language: Psychological and Biological Aspects*. Cambridge University Press.

I.J. Good. 1953. The population frequencies of species and the estimation of population parameters. *Biometrika*, 40(3 and 4):237–264.

G. Grefenstette. 1992. Finding semantic similarity in raw text: the Deese antonyms. In *Fall Symposium on Probability and Natural Language*. AAAI, October 23-25.

H.P. Grice. 1975. Logic and conversation. In P. Cole and J. Morgan, editors, *Syntax and Semantics III - Speech Acts*, pages 41–58. Academic Press, New York.

Jane Grimshaw. 1979. Complement selection and the lexicon. *Linguistic Inquiry*, 10(2):279–326.

Jane Grimshaw. 1990. *Argument Structure*. MIT Press, Cambridge, MA.

Jane Grimshaw. 1993. The least lexicon. Colloquium of the Institute for Research in Cognitive Science, University of Pennsylvania.

Ralph Grishman and John Sterling. 1992. Acquisition of selectional patterns. In *Proceedings of the Fourteenth International Conference on Computational Linguistics (COLING '92)*, pages 658–664, Nantes, France, July.

Ralph Grishman and John Sterling. 1993. Smoothing of automatically generated selectional constraints. In Madeleine Bates, editor, *ARPA Workshop on Human Language Technology*, March.

Jess Gropen. 1992. Constraints on a theory of verb learning: insights from polysemy. Presented at the 17th Boston University Conference on Language Development.

Jess Gropen. 1993. Participant types and the acquisition of verb polysemy. ms., May.

Stephen José Hanson. 1990. Conceptual clustering and categorization: bridging the gap between induction and causal models. In Yves Kodratoff and Ryszard S. Michalski, editors, *Machine learning : an artificial intelligence approach*, volume 3, pages 235–268. Morgan Kaufmann.

Marti A. Hearst and Kenneth W. Church. (in preparation). An investigation of the use of lexical associations for prepositional phrase attachment.

Marti Hearst and Hinrich Schütze. 1993. Customizing a lexicon to better suit a computational task. In *Proceedings of the ACL SIGLEX Workshop*, Columbus, Ohio, June.

D. Hindle and M. Rooth. 1991. Structural ambiguity and lexical relations. In *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics*, June. Berkeley, California.

D. Hindle and M. Rooth. 1993. Structural ambiguity and lexical relations. *Computational Linguistics*, 19(1):103–120, June.

Donald Hindle. 1983. User manual for Fidditch, a deterministic parser. Technical memorandum 7590-142, Naval Research Laboratory.

D. Hindle. 1990. Noun classification from predicate-argument structures. In *Proceedings of the 28th Annual Meeting of the Assocation of Computational Linguistics, Pittsburgh, Penna.*, pages 268–275.

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

Paul Hopper and Sandra Thompson. 1980. Transitivity in grammar and discourse. *Language*, 56(2):251–299.

Laurence Horn. 1989. *A Natural History of Negation*. University of Chicago Press.

Ray Jackendoff. 1983. *Semantics and Cognition*. Current Studies in Linguistics Series. The MIT Press.

Ray Jackendoff. 1990. *Semantic Structures*. Current Studies in Linguistics Series. The MIT Press.

Anil K. Jain and Richard C. Dubes. 1988. *Algorithms for Clustering Data*. Prentice Hall.

Frederick Jelinek and Robert L. Mercer. 1980. Interpolated estimation of Markov source parameters from sparse data. In *Proceedings of the Workshop on Pattern Recognition in Practice*, Amsterdam, The Netherlands: North-Holland, May.

Karen Jensen and Jean-Louis Binot. 1987. Disambiguating prepositional phrase attachments by using on-line dictionary definitions. *Computational Linguistics*, 13(3):251–260.

P. N. Johnson-Laird. 1983. *Mental models : towards a cognitive science of language, inference, and consciousness*. Harvard University Press.

Michael I. Jordan. 1986. Serial order: A parallel distributed processing approach. Technical Report ICS-8604, Institute for Cognitive Science, University of California at San Diego, La Jolla, CA.

P. Jusczyk, K. Hirsh-Pasek, D. Kemler Nelson, L. Kennedy, A. Woodward, and J. Piwoz. 1992. Perception of acoustic correlates of major phrasal units by young infants. *Cognitive Psychology*, 24:252–293.

Hans Kamp and Barbara Partee. in progress. Prototype theory and compositionality. ms.

Shyam Kapur. 1992. *Computational Learning of Languages*. Ph.D. thesis, Cornell University. Also appears as Cornell CS Technical Report 91-1234, September 1991.

Lauri Karttunen and Stanley Peters. 1979. Conventional implicature. *Syntax and Semantics*, 11.

Dieter Kastovsky. 1980. Selectional restrictions and lexical solidarities. In Dieter Kastovsky, editor, *Perspektiven der lexikalischen Semantik*, pages 70–92. Bonn: Bouvier Verlag Herbert Grundmann.

J. J. Katz and J. A. Fodor. 1964. The structure of a semantic theory. In J. A. Fodor and J. J. Katz, editors, *The Structure of Language*, chapter 19, pages 479–518. Prentice Hall.

J. J. Katz. 1970. Interpretative semantics vs. generative semantics. *Foundations of Language*, 6:220–259.

Slava M. Katz. 1987. Estimation of probabilities from sparse data for the language model component of a speech recognizer. *IEEE Transactions on Acoustics, Speech and Signal Processing*, ASSP-35(3):400–401, March.

D. Kemler Nelson, K. Hirsh-Pasek, P. Jusczyk, and K. Cassidy. 1989. How the prosodic cues in motherese might assist language learning. *Journal of Child Language*, 16:55–68.

D. Kemler Nelson. 1989. Developmental trends in infants' sensitivity to prosodic cues correlated with linguistic units. Presented at the biennial meeting of the Society for Research in Child Development, Kansas City, April. manuscript.

A. I. Khinchin. 1957. *Mathematical Foundations of Information Theory*. New York: Dover Publications. Translated by R. A. Silverman and M. D. Friedman.

John Kimball. 1973. Seven principles of surface structure parsing in natural language. *Cognition*, 2:15–47.

Kevin Knight. 1993. Building a large ontology for machine translation. In Madeleine Bates, editor, *ARPA Workshop on Human Language Technology*, March.

S. Kullback and R. A. Leibler. 1951. On information and sufficiency. *Annals of Mathematical Statistics*, 22:79–86.

S. Kurohashi and M. Nagao. 1992. Dynamic programming method for analyzing conjunctive structures in Japanese. In *Proceedings of COLING-92*, Nantes, France, August.

Barbara Landau and Lila Gleitman. 1985. *Language and Experience*. Harvard University Press, Cambridge, MA.

Anne Lederer and Michael Kelly. 1991. Prosodic correlates to the adjunct/complement distinction in motherese. *Papers and Reports in Child Language Development*, 30.

Anne Lederer, Henry Gleitman, and Lila Gleitman. forthcoming. The syntactic contexts of maternal verb use. ms.

Anne Lederer. 1993. *Title?* Ph.D. thesis, University of Pennsylvania.

Adrienne Lehrer. 1970. Verbs and deletable objects. *Lingua*, 25:227–254.

D. Lenat, M. Prakash, and M. Shepherd. 1986. CYC: using common sense knowledge to overcome brittleness and knowledge acquisition bottlenecks. *AI Magazine*, VI:65–85.

Beth Levin. 1989. Towards a lexical organization of English verbs. Technical report, Dept. of Linguistics, Northwestern University, November.

J. Lyons. 1961. *A structural theory of semantics and its application to lexical sub-systems in the vocabulary of Plato*. Ph.D. thesis, University of Cambridge, England. Published as *Structural Semantics*, No. 20 of the Publications of the Philological Society, Oxford, 1963.

Maryellen MacDonald. in press. The interaction of lexical and syntactic ambiguity. *Journal of Memory and Language*.

Maryellen MacDonald. in revision. Probabilistic constraints and syntactic ambiguity resolution. ms.

Brian MacWhinney and Catherine Snow. 1985. The Child Language Data Exchange System. *Journal of Child Language*, 12.

Brian MacWhinney. 1991. *The CHILDES project : tools for analyzing talk.* Erlbaum.

David Magerman. 1993. Parsing as statistical pattern recognition. ms., May.

Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of English: the Penn Treebank. *Computational Linguistics*, 19:313–330.

Mitchell Marcus. 1980. *A Theory of Syntactic Recognition for Natural Language*. MIT Press, Cambridge, MA.

Gail Mauner, Michael Tanenhaus, and Greg Carlson. 1992. Getting something for nothing: implicit arguments in sentence processing. ms.

James McCawley. 1968. The role of semantics in a grammar. In Emmon Bach and Robert Harms, editors, *Universals in Linguistic Theory*, pages 124–169. Holt, Rinehart and Winston.

James D. McCawley. 1971. Interpretative semantics meets Frankenstein. *Foundations of Language*, 7:285–296.

Kathleen McKeown and Vasileios Hatzivassiloglou. 1993. Augmenting lexicons automatically: Clustering semantically related adjectives. In Madeleine Bates, editor, *ARPA Workshop on Human Language Technology*, March.

Ken McRae, Virginia de Sa, and Mark S. Seidenberg. 1992. The role of correlated properties in accessing conceptual memory. Submitted to *Cognitive Psychology*.

George A. Miller and Walter G. Charles. 1991. Contextual correlates of semantic similarity. *Language and Cognitive Processes*, 6(1):1–28.

George Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine Miller. 1990. Five papers on WordNet. CSL Report 43, Cognitive Science Laboratory, Princeton University, July.

George A. Miller. 1971. Empirical methods in the study of semantics. In D. Steinberg and L. Jakobovits, editors, *Semantics, an interdisciplinary reader in philosophy, linguistics, and psychology*, pages 569–585. Cambridge University Press.

George Miller. 1990a. Nouns in WordNet: A lexical inheritance system, July. CSL Report 43, Princeton University. Also appears in *International Journal of Lexicography*, 3(4), 1990.

George Miller. 1990b. Wordnet: An on-line lexical database. *International Journal of Lexicography*, 3(4). (Special Issue).

Roy C. Milton. 1964. An extended table of critical values for the Mann-Whitney (Wilcoxon) two-sample statistic. *Journal of the American Statistical Association*, 59:925–934.

A. Mittwoch. 1971. Idioms and Unspecified NP Deletion. *Linguistic Inquiry*, 2:255–259.

A. Mittwoch. 1982. On the difference between *eating* and *eating something*: activities versus accomplishments. *Linguistic Inquiry*, 13(1):113–122.

J. Morris and G. Hirst. 1991. Lexical cohesion computed by thesaural relations as an indicator of the structure of text. *Computational Linguistics*, 17(1):21–48.

K. Nelson. 1973. Structure and strategy in learning to talk. *Monographs of the Society for Research in Child Development*, 38(1-2). Serial no. 149.

Sergei Nirenburg and Victor Raskin. 1987. The subworld concept lexicon and the lexicon management system. *Computational Linguistics*, 13(3-4):276–289.

D. N. Osherson and E. E. Smith. 1981. On the adequacy of prototype theory as a theory of concepts. *Cognition*, 9:35–58.

Patrick Paroubek, Yves Schabes, and Aravind K. Joshi. 1992. XTAG – a graphical workbench for developing tree-adjoining grammars. In *Third Conference on Applied Natural Language Processing*, Trento, Italy.

Neil Pearlmutter and Maryellen MacDonald. 1993. Plausibility and syntactic ambiguity resolution. In *Proceedings of the 14th Annual Conference of the Cognitive Science Society*, pages 498–503, Hillsdale, NJ. Erlbaum.

Fernando Pereira, Naftali Tishby, and Lillian Lee. 1993. Distributional clustering of English words. In *Proceedings of ACL-93*, June.

David Pesetsky. 1982. *Paths and Categories*. Ph.D. thesis, Massachusetts Institute of Technology.

Steven Pinker. 1989. *Learnability and Cognition*. MIT Press, Cambridge, MA.

W.H. Press, B.P. Flannery, S.A. Teukolsky, and W.T. Vetterling. 1988. *Numerical Recipes in C*. Cambridge University Press.

James Pustejovsky. 1991. The generative lexicon. *Computational Linguistics*, 17(4).

J. R. Quinlan. 1990. Induction of decision trees. In Jude W. Shavlik and Thomas G. Dietterich, editors, *Readings in Machine Learning*. Morgan Kaufmann. Originally published in *Machine Learning* 1:81–106, 1986.

Philip Resnik and Marti Hearst. 1993. Syntactic ambiguity and conceptual relations. In Kenneth Church, editor, *Proceedings of the ACL Workshop on Very Large Corpora*, pages 58–64, June.

Philip Resnik. 1991. An investigation of lexical class acquisition using a recurrent neural network. ms., December.

Philip Resnik. 1992a. A class-based approach to lexical discovery. In *Proceedings of the 30th Annual Meeting of the Association for Computational Linguistics*, Newark, Delaware, June. (Student session).

Philip Resnik. 1992b. Probabilistic tree-adjoining grammar as a framework for statistical natural language processing. In *Proceedings of the Fourteenth International Conference on Computational Linguistics (COLING '92)*, Nantes, France, July.

Philip Resnik. 1992c. WordNet and distributional analysis: A class-based approach to lexical discovery. In *AAAI Workshop on Statistically-based NLP Techniques*, San Jose, California, July.

Philip Resnik. 1993. Semantic classes and syntactic ambiguity. ARPA Workshop on Human Language Technology, March. Princeton.

Sally Rice. 1988. Unlikely lexical entries. In *Proceedings of the Berkeley Linguistics Society*, pages 202–212.

Matthew Rispoli. 1992. Discourse and the acquisition of *eat*. *Journal of Child Language*.

Luigi Rizzi. 1986. Null objects in Italian and the theory of *pro*. *Linguistic Inquiry*, 17(3):501–557, Summer.

Thomas Roeper. 1987. Implicit arguments and the head complement relation. *Linguistic Inquiry*.

Eleanor Rosch, Carolyn Mervis, Wayne Gray, David Johnson, and Penny Boyes-Braem. 1976. Basic objects in natural categories. *Cognitive Psychology*, 8:382–439.

Roger Schank. 1986. Language and memory. In B. Grosz, K. Sparck Jones, and B. Webber, editors, *Readings in Natural Language Processing*, pages 171–192. Morgan Kaufmann. Originally appeared in *Cognitive Science* 4(3), pp. 243–284, 1980.

Hinrich Schütze. 1993. Part-of-speech induction from scratch. In *ACL-93*.

Hinrich Schütze. (to appear). Word space. In S. J. Hanson, J. D. Cowan, and C. L. Giles, editors, *Advances in Neural Information Processing Systems 5*, San Mateo CA. Morgan Kaufmann.

R. W. Schvaneveldt, F. T. Durso, and D. W. Dearholt. 1989. Network structures in proximity data. In G. Bower, editor, *The psychology of learning and motivation: advances in research and theory*, volume 24, pages 249–284. Academic Press, New York.

Satoshi Sekine, Sofia Ananiadou, Jeremy Carroll, and Jun'ichi Tsujii. 1992. Linguistic knowledge generator. In *Proceedings of COLING-92*, pages 560–566, Nantes, France, August.

Claude E. Shannon and Warren Weaver. 1949. *The Mathematical Theory of Communication*. University of Illinois Press.

John Sinclair (ed.). 1987. *Collins COBUILD English Language Dictionary*. Collins: London.

Jeffrey Mark Siskind. 1992. *Naive Physics, Event Perception, Lexical Semantics, and Language Acquisition*. Ph.D. thesis, MIT.

Jeffrey Mark Siskind. 1993a. Lexical acquisition as constraint satisfaction. ms.

Jeffrey Mark Siskind. 1993b. Solving a lexical acquisition task via an encoding as a propositional satisfiability problem. Poster at the 6th Annual CUNY Sentence Processing Conference, March.

Frank Smadja. 1991. Macrocoding the lexicon with co-occurrence knowledge. In Uri Zernik, editor, *Lexical Acquisition: Exploiting On-Line Resources to Build a Lexicon*. Lawrence Erlbaum Associates.

E. E. Smith and D. L. Medin. 1981. *Categories and Concepts*. Harvard University Press.

Edward Smith and Daniel Osherson. 1988. Compositionality and typicality. In S. Schiffer and S. Steele, editors, *Cognition and Representation*, chapter 3, pages 37–52. Boulder, Colorado: Westview Press.

Paul Smolensky, Geraldine Legendre, and Yoshiro Miyata. 1992. Principles for an integrated connectionist/symbolic theory of higher cognition. Report CU-CS-600-92, Computer Science Dept., Univ. of Colorado at Boulder, July.

Paul Smolensky. 1988. On the proper treatment of connectionism. *Behavioral and Brain Sciences*, 11:1–74.

Padhraic Smyth and Rodney Goodman. 1992. An information theoretic approach to rule induction from databases. *IEEE Transactions on Knowledge and Data Engineering*, 4(4):301–316, August.

Jeffrey L. Sokolov and Catherine E. Snow, editors. (to appear). *Handbook of Research in Language Development using CHILDES*. Erlbaum Associates.

Karen Sparck Jones. 1964. *Synonymy and Semantic Classification*. Ph.D. thesis, University of Cambridge, England. Published in the Edinburgh Information Technology Series (EDITS), Sidney Michaelson and Yorick Wilks (eds.), Edinburgh University Press: Edinburgh, Scotland, 1986.

Patrizia Tabossi, Michael Spivey-Knowlton, Ken McRae, and Michael Tanenhaus. (in press). Semantic effects on syntactic ambiguity resolution: evidence for a constraint-based resolution process. *Attention and Performance,* XV.

Michael Tanenhaus, Susan Garnsey, and Julie Boland. 1991. Combinatory lexical information and language comprehension. In G. Altmann, editor, *Cognitive Models of Speech Processing: Psycholinguistic and Computational Perspectives*, pages 383–408. MIT Press.

A. M. Treisman. 1965. Verbal responses and contextual constraints in language. *Journal of Verbal Learning and Verbal Behavior*, 4:118–128. Reprinted in Oldfield, R. and Marshall, J., eds., *Language*, Penguin, 1968.

John Trueswell, Michael Tanenhaus, and Susan Garnsey. 1993. Evidence for the immedate use of local semantic constraints in syntactic ambiguity resolution. Presented at the 6th Annual CUNY Sentence Processing Conference, March.

John Trueswell. 1993. *The Use of Verb-Based Subcategorization and Thematic Role Information in Sentence Processing*. Ph.D. thesis, University of Rochester.

B. van Fraasen. 1968. Presuppositions, implications, and self-reference. *Journal of Philosophy*, 65:136–152.

Joh van Rooij and Reinier Plomp. 1991. The effect of linguistic entropy on speech perception in noise in young and elderly listeners. *Journal of the Acoustical Society of America*, 90(6):2985–2991, December.

Paola Velardi, Maria Teresa Pazienza, and Michela Fasolo. 1991. How to encode semantic knowledge: a method for meaning representation and computer-aided acquisition. *Computational Linguistics*, 17(2):153–170.

Paola Velardi. 1991. Acquiring a semantic lexicon for natural language processing. In Uri Zernik, editor, *Lexical Acquisition: Exploiting On-Line Resources to Build a Lexicon*, pages 341–368. Erlbaum.

Zeno Vendler, 1967. *Verbs and Times*, pages 97–121. Cornell University Press.

James Waldo. 1979. A PTQ semantics for sortal incorrectness. In *Linguistics, Philosophy, and Montague Grammar*. University of Texas Press.

Ralph Weischedel, Marie Meteer, Richard Schwartz, and Jeff Palmucci. 1989. Coping with ambiguity and unknown words through probabilistic models. ms.

Ralph Weischedel, Damaris Ayuso, R. Bobrow, Sean Boisen, Robert Ingria, and Jeff Palmucci. 1991. Partial parsing: a report of work in progress. In *Proceedings of the Fourth DARPA Speech and Natural Language Workshop, February 1991*.

Ralph Weischedel. 1986. A new semantic computation while parsing: presupposition and entailment. In B. Grosz, K. Sparck Jones, and B. Webber, editors, *Readings in Natural Language Processing*, pages 313–326. Morgan Kaufmann. Originally appeared in C. Oh and D. Dineen, eds., *Syntax and Semantics II: Presupposition and Entailment*, pp. 155–182, Academic Press, 1979.

Sholom M. Weiss and Casimir A. Kulikowski. 1991. *Computer systems that learn: classification and prediction methods from statistics, neural nets, machine learning, and expert systems.* Morgan Kaufmann, San Mateo, CA.

Greg Whittemore, Kathleen Ferrara, and Hans Brunner. 1990. Empirical study of predictive powers of simple attachment schemes for post-modifier prepositional phrases. In *Proceedings of the 28th Annual Meeting of the Association for Computational Linguistics*, pages 23–30. Pittsburgh, Pennsylvania.

Yorick Wilks and Dan Fass. 1992. The preference semantics family. *Computers & Mathematics with Applications*, 23(2–5):205–221.

Yorick Wilks, Xiuming Huang, and Dan Fass. 1985. Syntax, preference and right attachment. In *IJCAI-85*, pages 779–784.

Yorick Wilks. 1986. An intelligent analyzer and understander of English. In B. Grosz, K. Sparck Jones, and B. Webber, editors, *Readings in Natural Language Processing*, pages 193–204. Morgan Kaufmann. Originally appeared in *CACM* 18(5), pp. 264–274, 1975.

William Woods and James Schmolze. 1991. The KL-ONE family. *Computers and Mathematics with Applications, Special Issue on Semantic Networks in Artificial Intelligence*. Also available as Technical Report TR-20-90, Aiken Computation Laboratory, Harvard University.

Anthony Woods, Paul Fletcher, and Arthur Hughes. 1986. *Statistics in Language Studies*. Cambridge Textbooks in Linguistics. Cambridge University Press: Cambridge, England.

Fei Xu and Steven Pinker. 1992. Weird past tense forms. Presented at the 17th Boston University Conference on Language Development.

David Yarowsky. 1992. Word-sense disambiguation using statistical models of Roget's categories trained on large corpora. In *Proceedings of COLING-92*, pages 454–460, Nantes, France, July.

David Yarowsky. 1993. One sense per collocation. DARPA Workshop on Human Language Technology, March. Princeton.

Name Index

Akmajian, A., 77
Allen, R., 15
Alshawi, H., 21
Altmann, G., 104
American Heritage Dictionary, 25, 27
Ananiadou, S., 74
Armstrong, S, 50
Ayuso, D., 3, 21, 72, 105, 123, 124, 126

Bahl, L., 10, 11
Bar-Hillel, Y., 54, 58, 132
Basili, R., 3, 74, 105, 125, 126, 131
Beckwith, R., 5, 22, 63
Bensch, P., 13, 14
Binot, J., 21, 105
Bobrow, R., 3, 105, 123, 124, 126
Boisen, S., 3, 105, 123, 124, 126
Boland, J., 69, 97
Boyes-Braem, P., 50, 98
Braden-Harder, L., 21
Breiman, L., 11
Bresnan, J., 105
Brill, E., 13
Brown, P., 11, 13, 15–19, 30, 97, 116
Browne, W., 92
Brunner, H., 105, 115, 123
Bulmer, M., 54
Byrd, R., 21

Calzolari, N., 21
Carlson, G., 69, 97
Carroll, J., 74
Cassidy, K., 98
Chang, J., 3, 124, 126
Charles, W., 12, 108, 144
Chodorow, M., 21
Chomsky, N., 34, 43, 44, 49, 63
Church, K., 5, 7, 9, 13, 28, 30, 31, 103, 114–116, 122, 123, 135, 136
Collins COBUILD Dictionary, 45, 81, 84, 93
Cote, S., 77, 80
Cover, T., 7, 56, 125
Crain, S., 104
Cruse, D., 24, 107
Cupples, L., 68, 69, 71

Dagan, I., 17, 122
Dahlgren, K., 105
de Sa, V., 26
Dearholt, D., 13
Deerwester, S., 16
Dell, G., 97
Della Pietra, V., 13, 15–19, 30, 116
de Souza, P., 11, 13, 15–19, 30, 116
Donlon, G., 21, 72
Dowty, D., 22, 24, 79
Drange, T., 35, 36, 46, 47, 49, 51, 52, 59, 67
Dubes, R., 12
Dumais, S., 16
Durso, F., 13

Elman, J., 15, 21
Essen, U., 11

Fasolo, M., 74, 125, 133
Fass, D., 72, 105
Fellbaum, C., 5, 22, 63, 77, 83, 94, 97
Ferrara, K., 105, 115, 123
Ferstl, E., 68
Fillmore, C., 79, 81, 84, 93
Fisher, C., 44, 100, 101
Flannery, B., 10
Fletcher, P., 119

Fodor, J.A., iv, 2, 35, 42–45, 49, 50, 52, 53, 61, 64, 72, 133
Fodor, J.D., iv, 2, 35, 42–44, 82
Ford, M., 105
Francis, W., 8, 62, 85
Frazier, L., 104
Friedman, J., 11
Furnas, G., 16

Gale, W., 5, 7, 9, 28, 115, 122, 123, 135, 136
Garnsey, S., 69, 97
Garrett, M., 50
Gazdar, G., 77
Gillette, J., 98, 100, 101
Gleitman, H., 44, 50, 98, 100, 101
Gleitman, L., 44, 50, 98, 100, 101
Good, I.J., 62, 134
Goodman, R., 58
Gray, W., 50, 98
Grefenstette, G., 14
Grice, H., 39, 40, 96, 97
Grimshaw, J., 37, 44, 78, 79, 97
Grishman, R., 3, 11, 12, 16, 17, 74, 105, 124, 126
Gropen, J., 63, 101
Gross, D., 5, 22, 63

Hall, G., 100, 101
Hanks, P., 5, 7, 9, 13, 115
Hanson, S., 97
Harshman, R., 16
Hatzivassiloglou, V., 14
Hearst, M., 29, 31, 116, 120, 134
Heny, F., 77
Hindle, D., 3, 5, 7–9, 14, 105, 114–116, 120, 121, 131
Hirsh-Pasek, K., 98, 100
Hirst, G., 107
Holmes, V., 68, 69, 71
Hopper, P., 93
Horn, L., 34, 36, 37, 39–42, 49–51, 54
Huang, X., 105
Hughes, A., 119

Ingria, R., 3, 105, 123, 124, 126

Itai, A., 122

Jackendoff, R., 24, 82
Jain, A., 12
Jelinek, F., 10, 11, 119
Jensen, K., 21, 105
Johnson, D., 50, 98
Johnson-Laird, P., iv, 2, 35, 47–50, 53, 54, 59
Jordan, M., 15
Joshi, A., 103
Jusczyk, P., 98, 100

Kamp, H., 50, 51
Kaplan, R., 105
Kapur, S., 2
Karttunen, L., 39
Kastovsky, D., 42, 45, 46, 49, 52
Katz, J., iv, 2, 35, 42–45, 49, 52, 53, 61, 64, 72, 133
Katz, S., 119, 124
Kegl, J., 77, 83, 94, 97
Kelly, M., 98, 100
Kemler Nelson, D., 98, 100
Kennedy, L., 100
Khinchin, A., 56
Kimball, J., 104
Klavans, J., 21
Klein, E., 77
Knight, K., 30
Kučera, H., 8, 62, 85
Kulikowski, C., 12, 13, 109
Kullback, S., iv
Kurohashi, S., 106

Lai, J., 13, 15–19, 30
Landau, B., 98, 100
Landauer, T., 16
Lederer, A., 84–86, 98, 101
Lee, L., 14, 17, 18, 20, 21, 28, 30, 31
Legendre, G., 15
Lehrer, A., 83, 90, 95
Leibler, R., iv
Lenat, D., 23
Levin, B., 76, 77, 80, 81

Luo, Y-F, 3, 124, 126
Lyons, J., 23, 26, 51

MacDonald, M., 2, 69, 97
MacLaughlin, D., 21, 72
MacWhinney, B., 62, 85, 86
Magerman, D., 13, 17, 124
Marcinkiewicz, M., 1, 62, 85
Marcus, M., 1, 13, 62, 85, 126
Mauner, G., 97
McCawley, J., iv, 2, 35, 44, 45, 49, 52, 53, 61, 66
McDowell, J., 105
McKeown, K., 14
McRae, K., 2, 26, 69, 72
Medin, D., 50
Mercer, R., 10, 11, 13, 15–19, 30, 31, 116, 119
Mervis, C., 50, 98
Meteer, M., 105, 123
Miller, G.A., iv, 2, 3, 5, 12, 19, 22–25, 63, 85, 108, 144
Milton, R., 86
Mittwoch, A., 92
Miyata, Y., 15
Morris, J., 107

Nagao, M., 106
Neff, M., 21
Nelson, K., 98
Nirenburg, S., 23

Olshen, R., 11
Osherson, D., 50

Palmucci, J., 3, 105, 123, 124, 126
Parkes, C., 50
Paroubek, P., 103
Partee, B., 50, 51
Patil, R., 103, 114
Pazienza, M., 3, 74, 105, 125, 126, 131, 133
Pearlmutter, N, 69, 97
Pereira, F., 14, 17, 18, 20, 21, 28, 30, 31
Pesetsky, D., 78
Peters, S., 24, 39, 79
Pinker, S., 2, 91, 99, 100
Piwoz, J., 100
Plomp, R., 55
Prakash, M., 23
Press, W., 10
Pullum, G., 77
Pustejovsky, J., 80

Quinlan, J.R., 11

Rakowitz, S., 100, 101
Ramshaw, L., 21, 72
Raskin, V., 23
Resnik, P., 1, 16, 21, 72, 120, 134
Rice, S., 83, 95, 97
Rispoli, M., 81
Rizzi, L., 79, 82, 94
Roeper, T., 77
Rooth, M., 3, 105, 114–116, 121, 131
Rosch, E., 50, 98

Sag, I., 77
Santorini, B., 1, 13, 62, 85
Savitch, W., 13, 14
Schütze, H., 16–19, 29, 31
Schabes, Y., 103
Schank, R., 72
Schmolze, J., 72
Schvaneveldt, R., 13
Schwartz, R., 105, 123
Seidenberg, M., 26
Sekine, S., 74
Shaked, V., 21, 72
Shannon, C., 55, 58, 132
Shepherd, M., 23
Sinclair, J., 45, 81, 84, 93
Siskind, J., 2, 100
Smadja, F., 5, 29
Smith, E., 50
Smolensky, P., 15
Smyth, P., 58
Snow, C., 62, 86
Sokolov, J., 62, 86
Sparck Jones, K., 5, 22–26, 31, 51
Spivey-Knowlton, M., 2, 69, 72
Steedman, M., 104

Steinbiss, V., 11
Sterling, J., 3, 11, 12, 16, 17, 74, 105, 124, 126
Stone, C., 11
Stowe, L., 68, 69, 71
Su, K-Y., 3, 124, 126

Tabossi, P., 2, 69, 72
Tanenhaus, M., 2, 69, 72, 97
Teukolsky, S., 10
Thomas, J., 7, 56, 125
Thompson, S., 93
Tishby, N., 14, 18, 20, 21, 28, 30, 31
Treisman, A., 55
Trueswell, J., 72, 97
Tsujii, J., 74

van Fraasen, B., 38
van Rooij, J., 55
Velardi, P., 3, 74, 105, 125, 126, 131, 133
Vendler, Z., 91
Vetterling, W., 10

Waldo, J., 38
Walker, E., 50
Wall, R., 24, 79
Wanner, E., 98, 100
Weaver, W., 55, 58, 132
Weischedel, R., 3, 21, 72, 105, 123, 124, 126
Weiss, S., 12, 13, 109
Whittemore, G., 105, 115, 123
Wilks, Y., 72, 73, 105
Woods, A., 119
Woods, W., 72
Woodward, A., 100

Xu, F., 2

Yarowsky, D., 5, 29, 122, 123, 133

Zadrozny, W., 21