THE UNIVERSITY of EDINBURGH

Title	Acquisition and modeling of lexical knowledge : a corpus-based investigation of systematic polysemy
Author	Lapata, Maria
Qualification	PhD
Year	2000

Digitisation Notes:

- pag6, 8, 14, 30, 48, 138, 236 missing from original numeration.

The Acquisition and Modeling of Lexical Knowledge

A Corpus-based Investigation of Systematic Polysemy

Maria Lapata

PhD
University of Edinburgh
2000

Abstract

This thesis deals with the acquisition and probabilistic modeling of lexical knowledge. A considerable body of work in lexical semantics concentrates on describing and representing systematic polysemy, i.e., the regular and predictable meaning alternations certain classes of words are subject to. Although the prevalence of the phenomenon has been long recognized, systematic empirical studies of regular polysemy are largely absent, both with respect to the acquisition of systematic polysemous lexical units and the disambiguation of their meaning.

The present thesis addresses both tasks. First, we use insights from linguistic theory to guide and structure the acquisition of systematically polysemous units from domain independent wide-coverage text. Second, we constrain ambiguity by developing a probabilistic framework which provides a ranking on the range of meanings for systematically polysemous words in the absence of discourse context.

We focus on meaning alternations with syntactic effects and exploit the correspondence between meaning and syntax to inform the acquisition process. The acquired information is useful for empirically testing and validating linguistic generalizations, extending their coverage and quantifying the degree to which they are productive. We acquire lexical semantic information automatically using partial parsing and a heuristic approach which exploits fixed correspondences between surface syntactic cues and lexical meaning. We demonstrate the generality of our proposal by applying it to verbs and their complements, adjective-noun combinations, and noun-noun compounds. For each phenomenon we rely on insights from linguistic theory: for verbs we exploit Levin's (1993) influential classification of verbs on the basis of their meaning and syntactic behavior; for compound nouns we make use of Levi's (1978) classification of semantic relations, and finally we look at Vendler's (1968) and Pustejovsky's (1995) generalizations about adjectival meaning.

We present a simple probabilistic model that uses the acquired distributions to select the dominant meaning from a set of meanings arising from syntactically related word combinations. Default meaning—the dominant meaning of polysemous words in the absence of explicit contextual information to the contrary—is modeled probabilistically in a Bayesian framework which combines observed linguistic dependencies (in the form of conditional probabilities) with linguistic generalizations (in the form of prior probabilities derived from classifications such as Levin 1993). Our studies explore a range of model properties: (a) its generality, (b) the representation of the phenomenon under consideration (i.e., the choice of the model variables), (c) the simplification of its parameter space through independence assumptions, and (d) the estimation of the model parameters. Our findings show that the model is general enough to account

for different types of lexical units (verbs and their complements, adjective-noun combinations, and noun-noun compounds) under varying assumptions about data requirements (sufficient versus sparse data) and meaning representations (corpus internal or corpus external).

Acknowledgements

I am grateful to my supervisors Alex Lascarides, Chris Brew and Steve Finch for continuous support and advice over the last three years. This thesis undoubtedly benefited from Alex's thoroughness, insightful comments, and penetrating criticism. Without Chris' contribution and keen eye for mathematical detail, this thesis would have been a lot worse. Steve also provided helpful advice and stimulating discussions. I would also like to thank my thesis examiners, John Carroll and Marc Moens, for providing comments and feedback on my research. Frank Keller and Scott McDonald gave me valuable comments and criticism with regard to almost everything in this thesis. Frank read thesis chapters several times and was always keen to point out flaws and inconsistencies. Scott read them not as many times as Frank but provided as many comments. I am grateful to them for their help and friendship. The work reported in this thesis has also benefited from discussions with Claire Cardie, Claire Grover, Mark Light, Katja Markert, Raymond Mooney, Massimo Poesio, Mark Steedman, Simone Teufel, and Janyce Wiebe.

I received helpful comments and suggestions when presenting my work at the AAAI-99 Doctoral Consortium and at the following conferences: ACL-99, EACL-99, and AAAI-00. Finally, I would like to thank the subjects and annotators who made the experiments reported in this thesis possible. The financial support of ESRC, the Lilian Voudouri Foundation, and the Alexander S. Onassis Foundation are gratefully acknowledged.

Declaration

I declare that this thesis was composed by myself and that the work contained therein is my own, except where explicitly stated otherwise in the text.

(*Ṁaria Lapata*)

Contents

1 Introduction **15**

1.1 Central Claims 15

1.2 Motivation 17

1.2.1 Regular Polysemy 17

1.2.2 Modeling Ambiguity 22

1.2.3 Surface Cueing 26

1.3 Overview of the Thesis 27

1.4 Published Work 29

2 Methodology **31**

2.1 Corpora 31

2.1.1 The British National Corpus 31

2.1.2 The Penn Treebank Corpus 32

2.2 Lexical Resources 32

2.2.1 The WordNet Lexical Database 33

2.2.2 Roget's thesaurus 34

2.2.3 The CELEX Lexical Database 34

2.2.4 The NOMLEX Lexicon 35

2.2.5 The COMLEX Lexicon 35

2.3 Parsing 35

2.3.1 The Gsearch Corpus Query System 36

2.3.2 Partial Parsing via Finite-state Cascades 36

2.4 Probabilistic Modeling 38

2.5 Evaluation 44

2.5.1 The Kappa Statistic 44

2.5.2 Magnitude Estimation 45

2.6 Summary 46

3 Diathesis Alternations **49**
3.1 Introduction . . . 49
3.2 Experiment 1: The Dative and Benefactive Alternations . . . 52
3.2.1 Introduction . . . 52
3.2.2 Method . . . 54
3.2.3 Results . . . 69
3.2.4 Discussion . . . 76
3.3 Experiment 2: The Conative Alternation . . . 77
3.3.1 Introduction . . . 77
3.3.2 Method . . . 77
3.3.3 Results . . . 84
3.3.4 Discussion . . . 89
3.4 Experiment 3: The Possessor Object Alternation . . . 90
3.4.1 Introduction . . . 90
3.4.2 Method . . . 91
3.4.3 Results . . . 94
3.4.4 Discussion . . . 98
3.5 General Discussion . . . 98
3.6 Experiment 4: Validation . . . 100
3.6.1 Introduction . . . 100
3.6.2 Method . . . 101
3.6.3 Results . . . 102
3.6.4 Discussion . . . 103
3.7 Related Work . . . 104
3.8 Summary . . . 109

4 A Probabilistic Model of Verb Class Ambiguity **111**
4.1 Introduction . . . 111
4.2 The Model . . . 116
4.3 Experiment 5: Using Subcategorization to Resolve Verb Class Ambiguity . . . 124
4.3.1 Method . . . 124
4.3.2 Results . . . 124
4.3.3 Discussion . . . 125
4.4 Experiment 6: Using Corpus Distributions to Derive Verb Class Preferences . . . 126
4.4.1 Method . . . 126
4.4.2 Results . . . 127
4.4.3 Discussion . . . 132
4.5 General Discussion . . . 133

4.6 Related Work . . . 134
4.7 Summary . . . 136

5 A Probabilistic Model of Adjective-Noun Ambiguity 139
5.1 Introduction . . . 139
5.2 The Model . . . 143
5.2.1 Formalization of Adjective-Noun Polysemy . . . 143
5.2.2 Parameter Estimation . . . 146
5.3 Experiment 7: Comparison against the Literature . . . 149
5.3.1 Method . . . 149
5.3.2 Results . . . 150
5.3.3 Discussion . . . 152
5.4 Experiment 8: Comparison against Human Judgments . . . 153
5.4.1 Method . . . 153
5.4.2 Results . . . 158
5.4.3 Discussion . . . 163
5.5 Experiment 9: Comparison against Naive Baseline . . . 165
5.5.1 Naive Baseline Model . . . 165
5.5.2 Method . . . 166
5.5.3 Results . . . 166
5.5.4 Discussion . . . 167
5.6 General Discussion . . . 167
5.7 Related Work . . . 169
5.8 Summary . . . 172

6 Compound Nouns 173
6.1 Introduction . . . 173
6.2 Experiment 10: Compound Noun Extraction . . . 177
6.2.1 Method . . . 177
6.2.2 Results . . . 178
6.3 Statistical scores . . . 180
6.4 Experiment 11: Evaluation of Statistical Scores (CoocF ≥ 1) . . . 182
6.4.1 Method . . . 182
6.4.2 Results . . . 183
6.5 Experiment 12: Evaluation of Statistical Scores (CoocF > 1) . . . 185
6.5.1 Method . . . 185
6.5.2 Results . . . 185
6.6 Experiment 13: Evaluation of Statistical Scores (CoocF ≥ 5) . . . 187
6.6.1 Method . . . 187

6.6.2 Results . . . 187
6.7 Experiment 14: Hapaxes . . . 187
6.7.1 Method . . . 187
6.7.2 Results . . . 189
6.8 Discussion . . . 190
6.9 Experiment 15: Decision Tree Learning . . . 191
6.9.1 Features for Discovering Compounds . . . 191
6.9.2 Agreement . . . 196
6.9.3 Method . . . 197
6.9.4 Results . . . 197
6.10 Discussion . . . 203
6.11 Related Work . . . 205
6.12 Summary . . . 207

7 A Probabilistic Model of Nominalizations 209
7.1 Introduction . . . 209
7.2 The Model . . . 213
7.2.1 Parameter Estimation . . . 214
7.2.2 Smoothing . . . 216
7.2.3 The Algorithm . . . 220
7.2.4 Agreement . . . 221
7.3 Experiment 16: Smoothing Variants . . . 222
7.3.1 Method . . . 222
7.3.2 Results . . . 223
7.4 Experiment 17: Decision Tree Learning . . . 227
7.4.1 Method . . . 227
7.4.2 Results . . . 227
7.5 Discussion . . . 229
7.6 Related Work . . . 232
7.7 Summary . . . 235

8 Conclusions 237
8.1 Main Findings . . . 237
8.2 Issues for Further Research . . . 239
8.2.1 Further Modeling and Acquisition Studies . . . 239
8.2.2 Methodological Issues . . . 241
8.2.3 The Lexicon . . . 242
8.2.4 Word Sense Disambiguation . . . 243
8.2.5 Semantic Defaults and Intuitions . . . 243

A Annotation Guidelines **245**

A.1 Verb Class Ambiguity 245

A.2 Acquisition of Compound Nouns 273

A.3 Interpretation of Nominalizations 274

B Instructions **277**

C Materials **281**

Bibliography **285**

Index of Citations **301**

Chapter 1

Introduction

This thesis is about the use of probabilistic methods for the discovery of linguistic knowledge. More specifically, it uses corpora as the primary resource for the acquisition of word meaning and proposes a simple probabilistic model for selecting the dominant meaning from a set of meanings arising from syntactically related word combinations. This chapter presents the motivation for corpus-based acquisition and modeling of word meaning. It also summarizes the central claims put forward in this thesis and gives an overview of its structure.

1.1. Central Claims

The importance of the lexicon has been widely acknowledged for a variety of Natural Language Processing (NLP) tasks (Boguraev and Briscoe 1989; Boguraev and Pustejovsky 1995b). Several data-driven approaches to natural language have focused on the identification, extraction, and encoding of lexical information in a corpus. Methods have been developed for the acquisition of, for instance, parts of speech (e.g., Brill 1993), noun compounds (e.g., Daille 1996), collocations (e.g., Smadja 1991), support verbs (e.g., Grefenstette and Teufel 1995), subcategorization frames (e.g., Manning 1993), phrase structure rules (e.g., Finch 1993), selectional restrictions (e.g., Resnik 1993), antonyms (e.g., Justeson and Katz 1995a), and word senses (e.g., Schütze 1998).

Linguistic information (taxonomic, syntactic, and functional) has been either implicitly or explicitly used for the automatic or semi-automatic discovery of lexical knowledge. Corpora annotated with part-of-speech information, machine readable dictionaries, taxonomies such as WordNet and Roget's thesaurus, as well as corpus-induced grammatico-syntactic relations (e.g., modification, subcategorization) have been exploited for a variety of acquisition tasks (Aone and McKee 1995; Grefenstette 1994; Hindle 1990; Lesk 1986; Poznański and Sanfilippo 1995; Resnik 1993; Rooth, Riezler, Prescher, Carroll, and Beil 1999; Yarowsky 1995).

Despite the use of linguistic information as a means to guide or constrain the acqui-

sition process, little attention has been paid to linguistic generalizations about the meaning of words and phrases. For instance, most work concerned with the automatic induction of word senses from corpora does not pay heed to what linguistic theory has to say about systematic polysemy. Despite the methodological differences of the various word sense disambiguation proposals (i.e., supervised versus unsupervised, thesaurus-based versus dictionary-based), most approaches concentrate on the meanings of individual words (e.g., the word *date* may mean "day of the month", "appointment", or "fruit") and largely ignore cases of regular polysemy that arise from the systematic relationship between meaning and syntax.

The aim of the present thesis is to investigate how the meaning of word combinations (as opposed to individual words) can be induced from corpora using insights from linguistic theory to guide the acquisition process. We focus on systematic meaning alternations that have syntactic effects and use the corpus to identify their likelihood. We exploit the acquired information in two complementary tasks. First, we show how corpus data can provide a rich resource for testing, quantifying, and acquiring linguistic generalizations. Second, we exploit the corpus-derived frequencies in a probabilistic framework which constrains ambiguity by placing a preference ordering on the space of possible meanings. Unlike word sense disambiguation approaches, we do not aim at deriving the right meaning in a particular discourse context. Instead, we derive the most dominant (i.e., likely) meaning out of a set of meanings without taking discourse influences into account. In this way, we determine for a given word combination its most likely meaning overall (i.e., across the corpus) instead of focusing on the meaning of individual corpus tokens. This approach is complementary to linguistic theory and word sense disambiguation: it provides a general methodology for quantifying linguistic generalizations and discovering meanings for polysemous lexical units and a probabilistic framework for the ranking of alternative interpretations; the acquired meanings and their ranking can be further exploited by word sense disambiguation methods which generally use static sense inventories and lack a priori access to meaning frequencies.

This thesis puts forward four main claims. The first claim is that linguistic theory can contribute to the acquisition of linguistic knowledge from corpora. Text corpora are primary sources of information about language use and a theoretical framework can effectively guide and structure the acquisition process as well as the interpretation of the data. We demonstrate this by conducting a series of experiments that discover information pertaining to linguistic generalizations.

The second main claim is that corpus data can contribute to linguistic theory. Information mined from the corpus can be used to test, quantify, and extend linguistic theory by taking into account realistic examples which do not constitute idealizations of linguistic phenomena. A potential benefit of this approach is the expansion of the empirical base of linguistics and an increase in the predictive power of linguistic theory. Our findings indicate that lexical information gleaned from corpora contributes to the discovery of new linguistic facts that have eluded

linguists' introspection.

The third central claim is that a simple probabilistic model can be devised that uses the acquired distributions of lexical information to select the dominant meaning from a set of meanings arising from syntactically related word combinations. The dominant meaning in the absence of discourse context is modeled probabilistically in a Bayesian framework in which distributional information (in the form of conditional probabilities) is combined with linguistic generalizations (in the form of prior probabilities). We present a model that derives salient meanings for verbs and their complements, noun-noun compounds, and adjective-noun combinations.

The fourth claim concerns the methodology employed throughout this thesis. We show that a shallow approach which combines partial parsing and linguistic heuristics performs reasonably well at acquiring lexical information. Taking into account linguistic insights, our approach exploits consistent correspondences between surface syntactic and morphological cues and lexical meaning. Our findings show that the approach can be used to discover novel data and to obtain quantitative information for the modeling of systematic polysemy.

In the following sections we discuss recent work in lexical semantics and natural language processing and then motivate our claims in this context (see Sections 1.2.1 and 1.2.2). We also provide an overview of the thesis (see Section 1.3).

1.2. Motivation

1.2.1. Regular Polysemy

A considerable body of work in lexical semantics concentrates on describing and representing regular polysemy, i.e., the regular and predictable sense alternations to which certain classes of words are subject. Apresjan (1973: 16) defines regular polysemy as follows:

> Polysemy of the word A with the meanings a_i and a_j is called regular if, in the given language, there exists at least one other word B with the meanings b_i and b_j, which are semantically distinguished from each other in exactly the same way as a_i and a_j and if a_i and b_i, a_j and b_j are non-synonymous.

Regular polysemy is a pervasive phenomenon observed with verbs, nouns, and adjectives. Consider the examples in (1.1) where the verb *dress* receives a causative interpretation when attested in the transitive frame (see (1.1a)) and an inchoative interpretation in the intransitive frame (see (1.1b)). The differences in meaning (accompanied by differences in argument structure) do not reflect an idiosyncratic property of *dress*. A wide range of English verbs (mostly denoting bodily care) behave like *dress*: *bathe*, *change*, *shave*, *shower*, and *wash*. A similar phenomenon is observed with *kick* and verbs which generally mean hit (e.g., *hit*, *beat*,

pound, etc.). In the transitive frame *kick* receives a telic interpretation (see (1.2a)), whereas in the prepositional variant *kick* is atelic; it describes an attempted action (see (1.2b)).

(1.1) a. David dressed the baby.
b. David dressed.

(1.2) a. He kicked the birds.
b. He kicked at the birds.

Regular polysemy is also prominent with nouns: count nouns can be converted to mass nouns (see example (1.3)), containers can be extended to their contents (see (1.4)), names of trees can be used to refer to their wood (see (1.5)). The phenomenon extends to noun combinations. Consider the compound *student criticism* in (1.6a) whose modifier *student* can be interpreted as the subject or object of the deverbal head *criticism*. Again this is not an idiosyncratic compound; there is class of compounds called nominalizations which behave like *student criticism* (see example (1.6b)). A different class of compounds is illustrated in (1.7) where the modifier indicates the location of the compound head. Adjectives like *fast* and *slow* are also systematically polysemous. In (1.8a) a *fast/slow plane* is "a plane that flies fast/slowly", whereas in (1.8b) a *fast/slow book* is "a book that reads fast/slowly".

(1.3) a. John bought three lambs/chickens/rabbits.
b. John eats lamb/chicken/rabbit.

(1.4) a. John broke the bottle/cup/glass.
b. John drank the bottle/cup/glass.

(1.5) a. Cattle graze beneath the oaks/mahoganies.
b. The ladder was built of solid oak/mahogany.

(1.6) a. I appreciate student criticism.
b. Peter likes student administration.

(1.7) a. The number of road accidents has increased.
b. The theatre orchestra was founded in 1944.

(1.8) a. fast/slow plane
b. fast/slow book

Accounts of regular polysemy in the lexical semantics literature are either (a) descriptive, i.e., they take the form of linguistic classifications where words are categorized on the basis of shared components of meaning or (b) representational, i.e., they are concerned with the organization and representation of lexical semantic information. For example, Levin (1993) observes that verbs have in common a range of properties concerning the expression and interpretation of their arguments, as well as the extended meanings they can manifest. On the basis of their common syntactic and semantic properties, Levin groups verbs into classes and the characteristics of a given class can be used in principle to predict additional members. Levi

(1978) and Warren (1978) among others have proposed classifications for the interpretation of compound nouns. These classifications aim to specify the set of semantic relations that hold between a compound head and its modifier. For example, the underlying relationship between *road* and *accident* in *road accident* is IN (see (1.7a)), whereas a different relationship, SUBJ or OBJ, relates *student* with *criticism* (see (1.6a)). On the basis of these relationships new compounds can be classified and consequently interpreted. Similar classifications have been proposed for adjectives (Levi 1978; Vendler 1968; Warren 1984).

Pustejovsky's (1995) representational theory of the generative lexicon offers an account of regular polysemy by proposing novel ways of organizing and representing lexical information. The generative lexicon involves four levels of representation: argument structure, event structure (identifying the particular event type for a verb or a phrase), qualia structure (specifying information relating to the meaning of nouns), and lexical inheritance structure (specifying the way in which a word is related to other words in the lexicon). Consider for instance the examples in (1.8). Rather than enumerating the various senses for *fast* and *slow*, Pustejovsky's theory derives their meaning by exploiting the semantics of the nouns they are in construction with. More specifically, the *qualia structure* of nouns specifies the possible events associated with different entities. For example, the telic (purpose) role of the qualia structure for *plane* has a value equivalent to *flying*. The telic role for *book* is *reading*. When *plane* is combined with *fast* it modifies the event associated with it (i.e., *flying*).

Copestake and Briscoe (1992) and Copestake (1992) model the general ideas of the generative lexicon in a linguistic framework based on typed feature structures like Head-Driven Phrase Structure Grammar (HPSG, Pollard and Sag 1994) which allows for the specification of default inheritance in the lexicon. Lexical regularities are encoded in a computationally efficient manner, since lexical inheritance ensures that information need only be stated once instead of many times for each separate word. Lascarides (1995) argues for a new version of default inheritance which allows lexical generalizations to persist as default beyond the lexicon. Lascarides (1995), Lascarides and Copestake (1998) and Copestake and Lascarides (1997) further propose that lexical defaults interact with pragmatic defaults in order to model how the discourse context triggers exceptions to lexical generalizations.

Lexical rules have been suggested as a way of expressing the productivity of regular polysemy (Briscoe and Copestake 1999; Ostler and Atkins 1992). Lexical rules are typically interpreted as conditional relationships between lexical entries. Assume that we wish to describe the fact that verbs like *dress* in (1.1) may receive a causative/inchoative interpretation. A hypothetical lexical rule would take intransitive verbs of bodily care as its input (corresponding to the inchoative reading) and derive transitive verbs as its output (corresponding to the causative reading). Lexical rules can be also used to capture the generalizations pertaining to compound nouns. For instance a lexical rule could combine a location denoting noun (e.g., *road*) and an event denoting noun (e.g., *accident*) to produce a compound noun denoting an event and its

location (e.g., *road accident*).

Although linguistic theory is well-suited for predicting possible meanings without exhaustively enumerating the various word senses, it is typically concerned with representing *all* possible meanings rather than the most *likely* ones. In linguistic classifications such as Levin (1993) or Levi (1978), different meanings are determined through class membership. Meaning determination through class formation reduces redundancy, without however constraining ambiguity. Consider again the examples in (1.1). Although Levin predicts that *dress* has two different meanings, no information is given with respect to the likelihood of either interpretation. The same is true for *student criticism* which can be interpreted as "criticism by students" or "criticism of students" (see (1.6a)) and *theatre orchestra* which may be interpreted as "the orchestra of the theatre" or "an orchestra for the theatre", besides "the orchestra in the theatre" (see (1.7b)). The adjective-noun combinations in (1.8) are particularly interesting, since the range of possible interpretations is virtually unlimited: a *fast plane* may be "a plane that flies, goes, runs, takes off, moves, arrives fast", etc.

The use of defaults captures the fact that polysemous words have dominant meanings. For example, the default telic role for *plane* is *flying*, however it can be overridden when discourse context triggers a different interpretation. Similarly, a subject interpretation can be considered as the default for *student criticism*. Defaults typically reflect linguistic intuitions and in most cases are not empirically derived, for example via inspection of corpora (although see Copestake 1995 and Briscoe and Copestake 1996 for exceptions). Even if corpora are taken into account, it is infeasible to manually derive defaults for all words displaying systematic polysemy. Another related issue is that sometimes different defaults must be specified for different classes, or for their individual members. As pointed out by Kilgariff (1992: 89–90), there are cases in which one meaning of a systematic relationship is most prominent for certain instantiations of it, while the other meaning is more prominent for other instantiations. Consider the tree/wood relationship illustrated in (1.5): intuitively the tree meaning is more salient for *oak*, whereas the wood meaning is more salient for *mahogany*. Likewise, the attempted action meaning is less prominent for the verb *kick* (see example (1.2b)). In sum, linguistic theory accounts for the different meanings of lexical units without taking their frequency into account. Even in cases where defaults are used, the likelihood of the meanings of a given word or word combination or the likelihood with which a certain word undergoes a potential sense alternation are not empirically derived, instead they are approximated by intuition.

Similar considerations apply to the use of lexical rules which typically encode exceptionless conditional generalizations concerning the sense alternations of a given word. Briscoe and Copestake (1999: 488) argue that lexical rules are semiproductive, since they are subject to blocking[1] (i.e., the existence of one word prevents the application of a productive rule that would give rise to a word having the same semantics as the already existing word), arbitrary

[1] Blocking is also known as "preemption by synonymy" (Clark and Clark 1979).

lexical gaps and varying degrees of conventionalization. Consider the case where a count noun becomes a mass noun. Although the process is fully productive for *lamb*, *chicken*, or *rabbit*, it is blocked for *cow* by the existence of a synonymous form, i.e., *beef*. Briscoe and Copestake further argue that the productivity of lexical rules should be estimated empirically by taking corpus evidence into account and sketch how this approach could be implemented.

Although linguistic generalizations capture systematic regularities in the meaning of words and phrases, they are neither exhaustive nor complete descriptions of linguistic behavior. Consider again the classification approach. It is not possible to enumerate all compounds that fall under the SUBJ, OBJ, or IN relations, especially since compounding is very productive and new compounds can be created on the spot to satisfy the speaker's communicative needs. Corpus data could, however, provide important information about the frequency of these relations. Likewise, the verb classes specified in Levin (1993) cannot possibly include all verbs undergoing the systematic meaning alternations; however, their general properties can be used to identify novel verbs in corpus data. Although the generative lexicon goes a long way towards avoiding redundancy in the representation of regular polysemy, it does not exhaustively specify the particular values the various lexicon structures may have. Consider the qualia structure for the nouns *plane* and *book*. There is a variety of events associated with these nouns other than *flying* and *reading* which would have to be known in order to account for the different meanings of *fast* and *slow*. Note that the qualia structure must be further specified for all nouns potentially modified by the adjectives *fast* or *slow*. The set of telic roles for *book* and *plane* has to be specified irrespectively of whether these are thought to be part of the word's lexical representation or general world knowledge about objects with properties similar to *books* or *planes*. Corpora provide quantitative information about words, be it lexical or related to world knowledge. Although the interaction of lexical with world knowledge is important for the interpretation of words embedded in specific discourse contexts (Asher and Lascarides 1995), we largely ignore the issue in this thesis and concentrate on the discovery of linguistic knowledge without attempting to characterize its status.

These considerations pose the following two problems for practical NLP applications (e.g., information extraction, machine translation) that could potentially make use of lexical semantic information: (a) the *range* of possible interpretations needs to be determined, given that linguistic classifications are not exhaustive and in several cases these interpretations are left unspecified and (b) semantic ambiguity needs to be constrained by providing a *ranking* of alternative interpretations. Although some initial work has been done to address the former problem by utilizing Pustejovsky's (1995) theory of the generative lexicon to acquire lexical semantic information from corpora (Anick and Pustejovsky 1990; Bergler 1991; Pustejovsky, Bergler, and Anick 1993), hardly any effort has gone into exploiting the quantitative information acquired from corpora in a probabilistic framework in order to control the proliferation of lexical semantic ambiguity.

The present thesis uses insights from linguistic theory to guide the acquisition of lexical semantic information. In particular, we focus on sense alternations with syntactic effects (see examples (1.1), (1.6), and (1.8)) and exploit the correspondence between meaning and syntax in order to structure the acquisition process. We use the acquired information to empirically test the validity of linguistic generalizations, to extend their coverage, and to quantify the degree to which the sense alternations are productive. The research in this thesis is also concerned with developing a probabilistic framework that provides the means to restrict the ambiguity inherent in linguistic theory by automatically (rather than manually) acquiring dominant meanings for phenomena subject to regular polysemy.

1.2.2. Modeling Ambiguity

Semantic ambiguity is one of the most pervasive phenomena in natural language. In English, for example, it has been estimated that over 40% of words have more than one meaning (Britton 1978). The prevalence of lexical ambiguity has been of concern not only to linguists but also to researchers interested in the computational treatment of language. The problem of resolving semantic ambiguity is known as word sense disambiguation and generally involves "the association of a given word in a text or discourse with a definition or meaning (sense) that is distinguishable from other meanings potentially attributable to that word" (Ide and Véronis 1998: 3). The task involves two steps: (a) determining the different senses for every word relevant to the text or discourse under consideration and (b) a procedure for assigning each occurrence of a word to its appropriate sense. The two steps are interrelated, as we shall see below, since the choice of the sense inventory for the disambiguation task has an impact on the type of procedure used for sense assignment.

A considerable amount of work on word sense disambiguation has been conducted within the context of Artificial Intelligence (Boguraev 1979, Hirst 1987, Edward and Stone 1975, Small 1980, Wilks 1975, see Jurafsky and Martin 2000 and Ide and Véronis 1998 for overviews). Most early approaches to word sense disambiguation relied on hand-coded representations and focused on a small number of ambiguous words. Recent work in word sense disambiguation has moved away from hand-crafted systems, in favor of exploiting large-scale resources such as dictionaries, thesauri, and corpora. The proposed approaches vary in terms of the sense inventory they employ, the machine learning method they adopt, the degree to which they exploit context in the disambiguation procedure, and the number and category of words they attempt to disambiguate.

Empirical approaches to word sense disambiguation are either supervised or unsupervised. Supervised methods presuppose the existence of a disambiguated corpus available for training (Gale, Church, and Yarowsky 1992; Yarowsky 1992). A learning system (e.g., a Bayesian classifier) is typically presented with a training set consisting of a set of corpus tokens where each occurrence of the ambiguous word is annotated with its contextually appropriate

sense. The system learns to classify new ambiguous tokens on the basis of the surrounding context. The approach depends critically on manual sense tagging—a process that has to be repeated for every word in every language and potentially differently for different domains. A variant of this approach uses bootstrapping to reduce the need for large amounts of sense-tagged data (Hearst 1991; Yarowsky 1995). A problem that typically affects corpus-based word sense disambiguation methods is *data sparseness*. Most methods rely on word co-occurrence statistics, but many of the possible co-occurrences are not observed even in large corpora.

Unsupervised methods avoid the use of sense-tagged data during training. Some unsupervised methods assume no prior inventory of word senses; instead, unlabeled corpus tokens are clustered into distinct groups which are seen as representing the different senses of a given word (Schütze 1998). Other unsupervised disambiguation algorithms rely on sense inventories provided by lexical resources (e.g., machine-readable dictionaries) and typically combine information found in the corpus with the lexical resource in order to assign the appropriate sense to a given word (Dini, Tomaso, and Segond 1998, Guthrie, Guthrie, Wilks, and Slator 1991, Lesk 1986, see Wilks, Slator, and Guthrie 1996 for an overview).

Approaches to word sense disambiguation further differ in the number, type, and granularity of senses they employ. Machine-readable dictionaries (MRDs) such as the Oxford Advanced Learner's Dictionary (OALD, Cowie 1989) and the Longman Dictionary of Contemporary English (LDOCE, Procter 1978) have been extensively used in word sense disambiguation research (see Ide and Véronis 1998 and Wilks et al. 1996 for overviews). Despite the fact that MRDs are primarily created for human use rather than computational tasks (Atkins and Levin 1991; Kilgariff 1992), they provide ample information about word senses and can be used for large-scale experiments. Taxonomic information has played an important role in word sense disambiguation research. Early work primarily has made use of Roget's thesaurus (Bryan 1973; Masterman 1957), whereas recent work has also explored the contribution of WordNet (Resnik 1997; Sussna 1993; Voorhees 1993).

Irrespectively of their methodological differences, almost all word sense disambiguation approaches rely on the contribution of context for identifying the meaning for an ambiguous word. Under the *bag-of-words* approach, context is defined as an unordered set of words surrounding the ambiguous word without taking their interrelations into account. Under the *relational* approach, context is considered in terms of some relation to the ambiguous word. The relation can be simply the distance between the context and the ambiguous word, or syntactic, semantic, collocational, and categorial, (i.e., part-of-speech). Most word sense disambiguation studies involve a small number of highly ambiguous words (see Ng and Lee 1996 and Wilks and Stevenson 1998 for exceptions) and focus primarily on nouns, although some work has been done on verbs (Ng and Lee 1996) and adjectives (Chao and Dyer 2000; Justeson and Katz 1995a).

A common problem in word sense disambiguation research is how to determine the

appropriateness of the sense inventory employed. As Ide and Véronis (1998: 23) point out, the sense divisions contained in dictionaries are often too fine-graded for the purposes of NLP-related tasks. Sense inventories with a fine degree of granularity contribute to the combinatorial explosion of ambiguity and increase the amount of data required by supervised methods to unrealistic proportions. Furthermore, sense distinctions contained in dictionaries are often too fine-grained for humans to distinguish between them (Kilgariff 1992). Theories of the lexicon such as Levin (1993), Levi (1978), and Pustejovsky (1995) model the regularities of the meaning of words and phrases at a higher level of granularity than what is traditionally assumed in word sense disambiguation tasks, yet very little attention has been paid to the potential of these theories in word sense disambiguation research. A notable exception is Buitelaar (1997), who proposes underspecified semantic tagging as an alternative to word sense disambiguation. Buitelaar's proposal capitalizes on Pustejovsky's (1995) theory of the generative lexicon, in which related senses are not enumerated, but are instead generated from rules that capture regularities in sense creation. Buitelaar (1997: 25) sees semantic tagging as "the first step in the interpretation process by assigning each lexical item *all* of its systematically related senses, from which further semantic processing steps can derive discourse dependent interpretations".

Linguistic generalizations can provide important cues for word sense disambiguation because they point to systematic correspondences between meaning and surface syntactic or morphological cues. For example, we can disambiguate the verb *dress* on the basis of its subcategorization frame: a transitive usage points to causative meaning (see (1.1a)), whereas an intransitive usage points to the inchoative meaning (see (1.1b)). Knowing that the head noun in *student administration* is a nominalization (see (1.6b)) helps constrain the number of potential interpretations for this compound. Another related issue is the disambiguation target. Work in word sense disambiguation typically focuses on the meanings of individual words. However, ambiguity arises from word combinations even when there is no ambiguity with respect to the individual words participating in the combination. Consider again the compound *student administration*. Even if we know that *student* refers to a pupil (instead of a person who studies, e.g., a student of English) and *administration* refers to a body of persons who administer (instead of the act or process of administering), *student administration* remains ambiguous with respect to the agent or the patient of the administration. A similar situation can be observed with adjective-noun combinations. For example, the word *plane* in the combination *slow plane* (see (1.8a)) may refer either to an aircraft, a shape (e.g., a two-dimensional plane) or a tool. Irrespectively of the specific sense *plane* may assume, the combination *slow plane* can be ambiguous, as it can be interpreted as "a plane or a tool that flies, goes, or works slowly".

This thesis focuses precisely on the ambiguity arising from syntactically related word combinations rather than individual words. More specifically, we concentrate on cases of systematic polysemy and propose a probabilistic model which determines the prevalent meaning out of a set of available meanings. Unlike word sense disambiguation, our model delivers the

most *likely* meaning across all discourse contexts, instead of the *appropriate* meaning in a specific discourse context. In other words, our model predicts that *slow book* most likely means "book that reads slowly", even in cases where discourse context favors an alternative interpretation (e.g., "book that is published slowly"). Our model's predictions will also be context invariant with respect to the interpretation of *road accident* and *student criticism*: the former will most likely mean "accident in the road" and the latter "criticism by students".

Our task is complementary to word sense disambiguation. Notice that most word sense disambiguation research either lacks a priori access to meaning frequencies (all meanings are equally likely in MRDs) or laboriously obtains this information through corpus annotation (see the discussion above about supervised methods). The sense frequencies derived by our model can be exploited to semi-automatically produce the data required by supervised word sense disambiguation methods. Furthermore, they can be directly combined with a word sense disambiguation procedure, either supervised or unsupervised. For example, the acquired sense frequencies can serve as the initial parameters for a procedure which iteratively learns the typical usages of words from a corpus. Alternatively, a process can be devised which detects conflicts between frequent meanings and discourse context in order to arrive at the appropriate sense for a given usage or which defaults to the most frequent meaning in the absence of explicit reliable discourse information to the contrary (see Copestake and Lascarides 1997 for a theoretical account).

Although distinct from word sense disambiguation, our task is methodologically related, as it also involves the choice of an appropriate sense inventory and a way of determining the most likely meaning. With respect to the former task, senses will be provided either by linguistic classifications (e.g., Levin's 1993 list of verb classes) or directly derived from the corpus (through parsing) by exploiting fixed correspondences between surface characteristics of language input and meaning. For the latter task, we propose a probabilistic model which provides a ranking of the set of possible interpretations in an unsupervised manner, without relying on the availability of a disambiguated corpus. Like word sense disambiguation approaches, we make use of context for the estimation of the model parameters. We only exploit relational information in close proximity to the disambiguation target, and more specifically, syntactic information in conjunction with shallow semantic information.

We demonstrate the generality of our proposal by applying it to verbs and their complements, noun-noun compounds, and adjective-noun combinations. For each phenomenon we rely on insights from linguistic theory: for verbs we exploit Levin's (1993) influential classification of verbs on the basis of their meaning and syntactic behavior; for compound nouns we make use of Levi's (1978) taxonomy of semantic relations, and finally we look at Vendler's (1968) and Pustejovsky's (1995) generalizations about adjectival meaning for analyzing adjective-noun combinations. Our approach does not ignore the data sparseness problem which is common in corpus-based word sense disambiguation. We explicitly address data

sparseness in our study of compound nouns and show that the lack of sufficient distributional evidence can be compensated by combining different smoothing methods which "recreate" the frequencies of syntactically related word combinations.

1.2.3. Surface Cueing

Recent work in word sense disambiguation and lexicon acquisition has abstained from the deep syntactic and semantic processing of corpora in favor of a more shallow approach that utilizes partial parsing and/or surface syntactic information (e.g., part-of-speech information). A common methodological approach, at least for corpus-based word sense disambiguation, is the annotation of the corpus with part-of-speech information and its segmentation into basic syntactic relations (e.g., noun phrases, prepositional phrases, verb groups). Syntactic information can be used to guide the disambiguation procedure (Yarowsky 1993) or be further combined with other information sources such as capitalization (Hearst 1991), taxonomies (Leacock, Chodorow, and Miller 1998), and morphological and collocational information (Bruce and Wiebe 1994).

Work in lexical acquisition has extensively used partial parsing and surface collocational information for a variety of tasks such as the acquisition of terminology (see Manning and Schütze 1999 for an overview), subcategorization frames (Brent 1993; Briscoe and Carroll 1997; Manning 1993), support verbs (Grefenstette and Teufel 1995), selectional preferences (Abney and Light 1999). Besides partial parsing and co-occurrence statistics, recent work in corpus-based acquisition has proposed ways of inducing lexical semantic information on the basis of what Light (1996) calls *surface cueing*. Surface cueing uses observable cues as indicators about meaning exploiting fixed correspondences between surface characteristics of language input and meaning. Light (1996) shows that morphological features of a word can systematically correspond to a word's meaning. For example, the occurrence of a verb in the progressive tense can be used as a cue for the non-stativeness of this verb (given that stative verbs cannot appear in the progressive tense, **John is loving himself*).

The surface cueing approach has been successfully applied to the discovery of hyponyms (Hearst 1992) and to a variety of semantic classification tasks. For instance, Hatzivassiloglou and McKeown (1995a) develop a method for selecting the semantically unmarked term out of a pair of antonymous adjectives by taking into account linguistic diagnostics for markedness (see Chapter 5 for details). Siegel (1999) uses a variety of linguistic indicators (e.g., aspect, tense, modification by temporal adverbs, etc.) to classify verbs into two aspectual classes (stativity and completedness). Merlo and Stevenson (1999) define features that tap directly into thematic role distinctions in order to classify verbs automatically into lexical semantic classes. All of these approaches exploit surface cue and lexical semantic correspondences only, without relying on information external to the corpus (e.g., annotation of lexical semantic information or lexical resources such as taxonomies).

In this thesis we use partial parsing for the extraction of syntactic information. We fur-

ther extend the surface cueing approach (Light 1996) to verbs and their complements, adjective-noun combinations, and noun-noun compounds. We explore the correspondences between syntax and semantics relying primarily on linguistic observations of these correspondences. Our findings show that lexical semantic information can be induced from corpora using approximations of linguistic insights. This points to the usefulness of a shallow approach for automatic lexicon acquisition. Although linguistically-informed, our approach trades off linguistic sophistication against model simplicity. (A simple model is a model with a small number of parameters). The research in this thesis also explores the empirical importance of these simplifying assumptions, i.e., whether they compromise performance.

1.3. Overview of the Thesis

This thesis is divided into three parts: a methodological part (Chapter 2), an acquisition part (Chapters 3 and 6) and a modeling part (Chapters 4, 5, and 7).

Chapter 2 spells out the methodology used for the experimental studies presented in the chapters to follow. We give an overview of the corpora, lexical resources, and parsing tools used for the extraction of lexical semantic information. We also outline our evaluation methodology and explain the fundamentals of the probabilistic model used throughout this thesis.

Chapter 3 focuses on the acquisition of lexical information from corpora. In particular, we attempt to validate the claim that lexicon acquisition and linguistic theory can be complementary. We start from Levin's (1993) theory of diathesis alternations and examine the extent to which they are empirically attested in corpus data. We acquire frames characteristic of diathesis alternations from the corpus using the surface cueing approach and show how the acquired subcategorization type and token frequencies can be used to empirically estimate the *productivity* of an alternation (i.e., whether the verbs listed in Levin are found to alternate in the corpus) and whether a verb or verb class are *typical* (i.e., representative) of the alternation, demonstrating how frequency data can be used to quantify linguistic theory. Furthermore, our acquisition studies discover novel alternating verbs which are not listed in Levin, indicating that corpus-based research can provide the means to empirically validate and complement theories of linguistic generalizations.

Chapter 4 introduces our first modeling study of lexical semantic ambiguity. We demonstrate that the ambiguity exhibited by Levin's (1993) classification of verb meanings can be constrained via a model that combines Levin's inventory of meanings and the corpus frequencies acquired in the previous chapter. We further show that a shallow approach that uses linguistic theory on a par with corpus frequencies is not only useful for quantifying linguistic generalizations but also for practical applications that could benefit from a probabilistic ranking of the set of possible interpretations (for example by selecting the most likely one). Our

results show that frequency distributions of subcategorization frames within and across classes can satisfactorily derive the most salient meaning for a polysemous verb in the absence of any explicit contextual or lexical semantic information to the contrary.

Chapter 5 further evaluates our probabilistic model and the surface cueing approach by looking at polysemous adjective-noun combinations. In contrast to the previous chapter, where we used a linguistic classification (i.e., Levin 1993) for the inventory of the meanings of verbs in relation to their arguments, we acquire the meanings of adjective-noun combinations from a large corpus and propose a probabilistic model which provides a ranking on the set of possible interpretations. We identify lexical semantic information automatically by exploiting the consistent correspondences between surface syntactic cues and lexical meaning. We evaluate our results against paraphrase judgments elicited experimentally from humans and show that the model's ranking of meanings correlates reliably with human intuitions: meanings that are found highly probable by the model are also rated as plausible by the subjects.

Chapter 6 reports a series of experiments on the acquisition of noun-noun compounds, a phenomenon that is extremely productive. Identifying compounds is challenging for the surface cueing approach given that any sequence of two nouns can be a potential compound in the right context. This means that not only frequent noun combinations are candidate compounds but also rare ones. Our findings demonstrate that a surface approach which looks for consecutive nouns in the corpus is sufficient only for frequent compounds. We present a series of experiments which show that rare compounds can also be identified using a combination of linguistic features (e.g., the frequency of the compound head, the likelihood of a noun as a modifier, the context surrounding the candidate compound). Our results indicate that the acquisition of compounds on the basis of the distributional properties of co-occurring nouns in the corpus yields satisfactory results even in cases where data is sparse.

Chapter 7 concentrates on a probabilistic model for the interpretation of nominalizations, a particular class of compound nouns, where the compound head is a deverbal noun and the modifier is its subject or object. As in Chapter 5, we acquire the inventory of the argument relations from the corpus. Although the degree of ambiguity exhibited by nominalizations is less than that exhibited by polysemous verbs and adjectives (note that in the case of nominalizations we have a binary choice between a subject or object relation) the task is less straightforward than it seems since the estimation of the model parameters runs into severe data sparseness problems. We show how this can be overcome by combining surface cueing (i.e., partial parsing), smoothing techniques, and domain independent taxonomic information (e.g., WordNet).

Chapter 8 summarizes the main findings of this thesis and outlines some suggestions for future research.

1.4. Published Work

Some of the material presented in this thesis has been published or is presently under review for publication; this applies to Chapter 1 (Lapata 1999c), Chapter 3 (Lapata 1999a), Chapter 4 (Lapata and Brew 1999), Chapter 5 (Lapata 2001), Chapter 6 (Lapata 1999b), and Chapter 7 (Lapata 2000).

Chapter 2

Methodology

This chapter focuses on methodological issues related to the acquisition of lexical semantic information from corpora and the probabilistic modeling of systematic polysemy. The experiments reported throughout this thesis will empirically test a hypothesis by automatically acquiring information from a corpus; this quantitative information will be exploited by a probabilistic model to make predictions about linguistic behavior. The model's predictions will be evaluated against humans. We will discuss and motivate our hypotheses in the following chapters. In this chapter we focus on the experimental methodology underlying the studies reported in this thesis. In particular, we discuss the corpora we use to extract lexical semantic information, the means we employ to obtain this information (i.e., shallow parsing), and the mathematical foundations of our probabilistic model. In some cases, we will use lexical resources external to the corpus to obtain morphological, syntactic or semantic information. We also briefly describe these resources in this chapter. Finally, we turn to evaluation and describe our methodology for assessing our experimental results.

2.1. Corpora

2.1.1. The British National Corpus

The corpus used for the majority of the research reported in this thesis is the British National Corpus (BNC, Burnard 1995). The BNC is a large, synchronic corpus, consisting of 90 million words of text and 10 million words of speech. The BNC is a balanced corpus, i.e., it was compiled so as to represent a wide range of present day British English. The written part includes samples from newspapers, magazines, books (both academic and fiction), letters, and school and university essays, among other kinds of text. The spoken part consists of spontaneous conversations, recorded from volunteers balanced by age, region, and social class. Other samples of spoken language are also included, ranging from business or government meetings to radio shows and phone-ins. The corpus represents many different styles and varieties, and is

not limited to any particular subject field, genre or register.

Furthermore, the BNC is annotated with part-of-speech information. The corpus was tagged automatically with CLAWS4, a probabilistic part-of-speech tagger, using a tagset of 61 distinct grammatical labels (Leech, Garside, and Bryant 1994). In the experiments reported throughout this thesis a part-of-speech annotated and lemmatized version of the BNC is used. The lemmatized version of the corpus was obtained using Karp, Schabes, Zaidel, and Egedi's (1992) morphological analyzer. This means that our frequency estimates will draw on counts obtained from lemmas rather than from word forms.

2.1.2. The Penn Treebank Corpus

Although the majority of our experiments are conducted on the BNC, the Penn Treebank corpus is used for a validation study reported in Chapter 3 (see Experiment 4). The Penn Treebank contains approximately one million words of written American English. In contrast to the BNC, it consists only of written text and is limited to a particular genre and subject field (financial newspaper articles taken from the Wall Street Journal). The Penn Treebank is not only annotated with part-of-speech but also with phrase structure information. The part-of-speech tagged version of the Penn Treebank was produced using a combination of automatic part-of-speech assignment (using PARTS, Church 1988) and manual correction. The part-of-speech tagset contains 36 distinct grammatical labels (see Marcus, Santorini, and Marcinkiewicz 1993 for details).

A similar methodology was employed for the syntactic annotation of the corpus. Fidditch (Hindle 1989), a partial parser (see Section 2.3 for details on partial parsing), was used to provide an initial parsed version of the corpus which was subsequently manually corrected. The syntactic annotation contains information about predicates and their arguments, adjuncts, null elements such as the subjects of infinitival constructions, coordinated, and coindexed constituents.

2.2. Lexical Resources

Although corpora will be our primary source of syntactic and lexical semantic information, in some instances this information will be complemented with knowledge external to the corpus. For example, in cases where the corpus does not provide enough evidence on its own because the phenomenon under investigation is sparse, a linguistic taxonomy will be used to recreate the missing evidence (see Chapters 6 and 7). In the work reported in this thesis we experimented with two taxonomies described below, WordNet (Miller, Beckwith, Fellbaum, Gross, and Miller 1990) and Roget's thesaurus.

External resources will be also consulted in cases where detailed morphological information is crucial for the modeling of systematic polysemy (see Chapter 7). We used two

Sense 1
bread, breadstuff, staff of life
⇒ baked good, baked goods
⇒ foodstuff, food product
⇒ food, nutrient
⇒ substance, matter
⇒ object, physical object
⇒ entity, something
Sense 2
shekels, gelt, dough, bread, dinero, lucre, loot, pelf, moolah, cabbage, kale
⇒ money
⇒ medium of exchange, monetary system
⇒ asset
⇒ possession

Figure 2.1: WordNet hypernyms of *bread*

resources for obtaining the latter information, CELEX (Burnage 1990) and NOMLEX (Macleod, Grishman, Meyers, Barrett, and Reeves 1998). In other instances lexical information acquired from the corpus will be compared against lexical resources (such as WordNet or CELEX) as a means to evaluate the results of an automatic procedure (see Chapters 3 and 6). Finally, we will make use of external resources in order to filter out erroneous information acquired from the corpus (see Chapter 3). In the following we give a brief overview of these resources, placing emphasis on information relevant for our experiments.

2.2.1. The WordNet Lexical Database

WordNet is an on-line lexical reference system whose design is inspired by current psycholinguistic theories of human lexical memory (Miller et al. 1990). The WordNet lexicon contains verbs, nouns, adjectives, and adverbs. Lexical information is organized in terms of word meanings, rather than word forms. Word meanings are represented by their synonyms or near synonyms. The different senses of a given word are denoted by the different synonym sets (henceforth synsets) the word belongs to. Synsets are further organized by semantic relations such as antonymy, hyponymy, and meronymy. Nouns are organized into an inheritance system defined by hypernymic (superordinate) relations. Nouns are not contained in a single hierarchy; instead they are partitioned according to a set of semantic primitives which are treated as the unique roots of separate hierarchies.

Semantic information for each noun is encoded in terms of its hyponyms, hypernyms, synsets, holonyms (i.e., concepts that have the noun as its parts), and meronyms (i.e., concepts that are parts of the noun). As an example the hypernyms of the noun *bread* are shown in Figure 2.1. Since *bread* has two senses it is defined by two synsets. The synset ⟨bread, breadstuff, staff of life⟩ corresponds to the "food" sense of *bread*, whereas the synset ⟨shekels, gelt,

dough, bread, dinero, lucre, loot, pelf, moolah, cabbage, kale⟩ corresponds to the "money" sense of *bread*. *Onion bread, raisin bread, quick bread,* and *rye bread* are hyponyms of *bread*. *Flour* is its meronym and *sandwich* is its holonym.

For the experiments reported in this thesis we concentrated only on the noun taxonomy (see Experiments 1–3 and 15–17) and the hypernymic/hyponymic relation. The WordNet noun taxonomy (version 1.6) contains 94,474 nouns (including compound nouns) and 4,795 distinct concepts.

2.2.2. Roget's thesaurus

Similarly to WordNet, Roget's thesaurus groups words (nouns, verbs, adjectives, and adverbs) into semantic categories. Roget's thesaurus, however, lacks WordNet's hierarchical organization. Semantic information about a given word is defined via its membership to one or more categories. Consider for example the category EXISTENCE. Members of this category are the nouns *entity, being, reality*, the verbs *exist, live, breath*, the adjectives *existent, substantial, real*, and the adverbs *actually, in fact, in reality*. A word can be a member of more than one category. For example, *being* is also a member of the category SUBSTANTIALITY and *breath* is member of the categories INSTANTANEITY, WIND, ANIMALITY, and FAINTNESS.

The experiments reported in this thesis used only the taxonomic information pertaining to nouns (see Experiments 15–17). Roget's thesaurus (version 1911) contains 20,446 nouns and 1,043 semantic categories.

2.2.3. The CELEX Lexical Database

CELEX (Burnage 1990) is a lexical database which has been developed for English, Dutch, and German. The database contains orthographical, phonological, morphological, and syntactic information for a large number of lexical items. For example, orthographic information about a given word specifies the number of spellings it has, whether it is American or British English, and its spelling frequency. Phonological information specifies the number of pronunciations for a given word, the status of pronunciation (i.e., whether it is primary or secondary), and its phonetic transcription. Morphological information specifies the inflectional features for a given word, the morphemes it consists of, its allomorphs, whether it is the product of a derivational process (e.g., *leakage* is derived either from the noun *leak* and the affix *-age* or from the verb *leak* and the affix *-age*), whether it is monomorphemic (e.g., *camel*) or morphologically complex (e.g., *breathalyse*). Nouns are further distinguished into compounds and non-compounds. The derivational processes leading to compound noun formation are also specified.

In this thesis we used the English version of CELEX for two tasks: (a) to compile a list of compound nouns (see the experiments in Chapter 6) and (b) to gather information about nouns derived from verbs (e.g., *adaptation* is derived from the verb *adapt*, *bet* is derived from

the verb *bet*, see Section 7.2.1.2 for details).

2.2.4. The NOMLEX Lexicon

NOMLEX (Macleod et al. 1998) is a dictionary of nominalizations (i.e., nouns derived from verbs). The dictionary goes beyond CELEX in that it not only specifies the underlying verb from which a given noun is derived but also the noun's complements. For example, the noun *destruction* is derived from the verb *destroy*, it can have a possessive noun phrase as its object (*Rome's destruction*) or subject (*Rome's destruction of Carthage*), and a noun or prepositional phrase as its object (*forest destruction*, *destruction of the forest*).

We concentrated only on the morphological information provided by NOMLEX to compile (in conjunction with CELEX) a lexicon of nominalizations (see Chapter 7 for details).

2.2.5. The COMLEX Lexicon

COMLEX (Grishman, Macleod, and Meyers 1994) is a subcategorization dictionary. The dictionary contains 38,000 entries and provides detailed subcategorization information for verbs, nouns, and adjectives. For each lexical item the dictionary entry specifies its part of speech, base form, and its complements, if it has any. For example, the lexical entry for the verb *abandon* informs us that it can occur with two arguments, an NP and a *to*-PP (e.g., *We abandoned our people to poverty*) or with just an NP (e.g., *He abandoned the study of law*).

We used only the subcategorization information pertaining to verbs (6,000 entries). In particular, on the basis of the COMLEX frames and their type frequencies we estimated the likelihood of a given frame and used this information to discard erroneous frames acquired from the BNC (see Experiments 1–3 in Chapter 3 for details).

2.3. Parsing

Throughout this thesis we will use corpora to extract information not about single words but about lexical relations. In Chapter 3 we study verbs and their subcategorization patterns, in Chapter 5 we look at adjective-noun modification, and in Chapter 6 we turn to compound nouns. We obtain this information focusing on shallow linguistic patterns as opposed to detailed syntactic analysis. More specifically, we either use regular expressions to search the corpus for basic syntactic relations (e.g., verb-complement relations, noun-noun relations) or recover these relations from the output of a partial (i.e., chunk) parser which produces a full but shallow analysis of unrestricted text. Although both approaches are ideal for the efficient analysis of text, the obtained information is in some cases imperfect. Emphasis is placed on recognizing basic syntactic units without attempting to resolve attachment ambiguities or to recover missing information (such as traces resulting from the movement of constituents). In

the following we describe Gsearch (Corley, Corley, Keller, Crocker, and Trewin 2001), a regular expression matcher, and Cass (Abney 1996), a partial parser which relies on cascades of finite-state automata to produce an analysis for unrestricted text.

2.3.1. The Gsearch Corpus Query System

Gsearch (Corley et al. 2001) is a tool which permits the search of arbitrary part-of-speech tagged corpora for shallow syntactic patterns based on a user-specified grammar and a syntactic query. It achieves this by combining a bottom-up chart parser with a regular expression matcher. The parser matches grammar terminals to corpus data fields using the regular expression matcher. The corpus fields may either be part of the corpus (e.g., words, lemmas, part-of-speech tags) or may be created using external linguistics resources (e.g., a lexical database such as WordNet). The parser attempts to build syntactic structure that matches the user-specified query.[1]

Depending on the grammar specification (i.e., recursive or not) Gsearch can be used as a full context-free parser or a chunk parser. Depending on the syntactic query, Gsearch can parse full sentences, identify syntactic relations (e.g., verb-object, adjective-noun) or even single words (e.g., all indefinite pronouns in the corpus). Gsearch outputs all corpus sentences containing substrings that match a given syntactic query. Given two possible parses that begin at the same point in the sentence, the parser chooses the longest match. If there are two possible parses that can be produced for the same substring, only one parse is returned. This means that if the number of ambiguous rules in the grammar is large, the correctness of the parsed output is not guaranteed.

The Gsearch system has been used for a number of corpus-based studies in psycholinguistics (Corley and Cuthbert 1997; Corley and Haywood 1999; Lapata, Keller, and Schulte im Walde 2001; Sturt, Pickering, and Crocker 1999), computational linguistics (Lapata, McDonald, and Keller 1999) and theoretical linguistics (Zamparelli 1998). We used Gsearch for the extraction of information relating to subcategorization frames (see Chapter 3 for details), adjective-noun combinations (see Chapter 5), and compound nouns (see Chapter 6).

2.3.2. Partial Parsing via Finite-state Cascades

Gsearch can efficiently provide quantitative information about syntactic relations without fully parsing the corpus. However, in some cases obtaining a full syntactic analysis, even if it simply consists of recognizing basic phrases can provide useful information, especially when we want to recover global information about the number of times a given noun is attested as the subject

[1] The parser's implementation is a modification of the standard chart parser algorithm (Earley 1970): the parser uses an oracle (which is calculated automatically from the grammar provided by the user) to discard any analysis that results in a constituent that cannot be a left-descendant of a current goal (i.e., user query). This optimization results in very fast processing of the corpus (see Corley et al. 2001 for details).

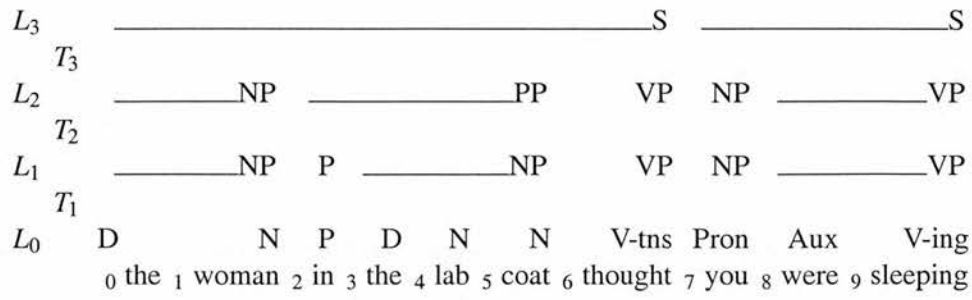

Figure 2.2: Parse tree represented as sequence of levels L_0–L_3

$$
\begin{array}{ll}
T_1: & \left\{ \begin{array}{l} \text{NP} \rightarrow \text{D? N* N} \\ \text{VP} \rightarrow \text{V-tns} \mid \text{Aux V-ing} \end{array} \right\} \\
T_2: & \{\text{PP} \rightarrow \text{P NP}\} \\
T_3: & \{\text{S} \rightarrow \text{PP* NP PP* VP PP*}\}
\end{array}
$$

Figure 2.3: Finite-state level recognizers T_1–T_3

or object of a given verb, or the number of times a verb is modified by an adverb, etc. We used Cass (Abney 1996) to obtain a shallow syntactic analysis of the corpus.

The main feature of Cass is the finite-state cascade technique. A finite-state cascade is a sequence of non-recursive levels: phrases at one level are built on phrases at the previous level without containing same level or higher-level phrases. Two levels of particular importance are *chunks* and *simplex clauses*. A chunk is the non-recursive core of intra-clausal constituents extending from the beginning of the constituent to its head, excluding post-head dependents (i.e., NP, VP, PP), whereas a simplex clause is a sequence of non-recursive clauses (Abney 1996). Cass recognizes chunks and simplex clauses using a regular expression grammar without attempting to resolve attachment ambiguities. Figure 2.2, taken from Abney (1996: 8), illustrates a parse tree represented as a sequence of levels.

Parsing is a series of finite transductions, each defined by a set of *patterns* consisting of a category and a regular expression (see Figure 2.3). These patterns are translated into finite-state automata (see Aho, Sethi, and Ullman 1986 for details on the translation process) and the union of all automata at a given finite transduction yields a single automaton, a *finite-state level recognizer* T_i. Consider for example the recognizer T_1 in Figure 2.3. It applies to level L_0 in Figure 2.2 and outputs the NP from position 0 to position 2. The recognizer restarts at position 2; since no transition is possible from position 2 to position 3, the recognizer outputs the category P and restarts from position 3, etc. Given two possible parses that begin at the same point in the sentence, the longest match is preferred (see Figure 2.2 where the recognizer outputs an NP from positions 3 to 6 instead of 3 to 5).

Parsing using finite-state cascades is not only fast, since unlike traditional parsers no global optimization takes place, but also fairly robust. Parsing proceeds by reliably recognizing

parts of syntactic structure leaving ambiguous constituents unattached rather than producing a global parse tree. When evaluated against human judgments the parser identified chunks with 87.9% precision and 87.1% recall (Abney 1996). The parser further achieved a per-word accuracy of 92.1% (where per-word accuracy includes the chunk category and chunk length identified correctly).

The parser comes with a large-scale grammar for English and a built-in tool that extracts predicate-argument tuples out of the parse trees that Cass produces. The parser has been used in a variety of studies in computational linguistics such as word sense disambiguation (Dagan, Lee, and Pereira 1999) and induction of selectional preferences (Abney and Light 1999). We used the parser's output in two studies which model the meaning of adjective-noun combinations (see Chapter 5) and nominalizations (see Chapter 7).

2.4. Probabilistic Modeling

This thesis is concerned with providing a probabilistic framework for systematic polysemy. More specifically, we are interested in finding the most likely meaning from the set of meanings allowed by a given word combination. Before discussing the properties of our probabilistic model consider the following problem, discussed in Russell and Norvig (1995) and Collins (1998), which bears little relation to linguistic meaning but illustrates nicely some of the desiderata underlying the specification of language-related models.

> Imagine a person has a new burglar alarm installed at home. It is fairly reliable at detecting a burglary, but also responds on occasion to minor earthquakes. This person also has two neighbors, John and Mary, who have promised to call her at work when they hear the alarm. John always calls when he hears the alarm, but sometimes confuses the telephone ringing with the alarm and calls then, too. Mary, on the other hand, likes rather loud music and sometimes misses the alarm altogether (Russell and Norvig 1995: 437).

Suppose that we want to estimate the probability of a burglary given that we know, for example, that Mary has called. Our first task is to choose the variables representing the problem, for example B (burglary event), E (earthquake event), A (alarm event), J (John calls), and M (Mary calls). Next, we need to specify a joint distribution for these variables. The ordering of the variables is crucial for the *compactness* (i.e., number of parameters) of the model. Suppose we choose the ordering $\langle B,E,A,J,M\rangle$. The joint probability $P(A,B,E,J,M)$ can be re-written using the chain rule as follows.

$$P(A,B,E,J,M) = P(B)\cdot P(E|B)\cdot P(A|B,E)\cdot P(J|B,E,A)\cdot P(M|B,E,A,J) \tag{2.1}$$

First, notice that the distribution in (2.1) is an approximation of the problem described above. For example, we choose not to model the fact that Mary likes loud music and as a result she sometimes misses the alarm or that John confuses the phone ringing with the alarm. Second, the distribution in (2.1) has five variables. Assuming that the variables are boolean (i.e., true or false), we have to specify $2^5 - 1 = 31$ parameters. We can, however, reduce the number of parameters and create a more compact model by making independence assumptions. For example, there is no causal relation between burglaries and earthquakes (see (2.2)). John will call if the alarm goes off independently of whether there is a burglary or an earthquake (see (2.3)). Similarly, Mary's calling is dependent on her hearing the alarm and independent of John calling or whether there is a burglary or an earthquake (see (2.4)). Reasoning about causality enables us to reduce the parameter space as follows:

$$P(E|B) = P(E) \tag{2.2}$$

$$P(J|A,E,B) = P(J|A) \tag{2.3}$$

$$P(M|A,E,B,J) = P(M|A) \tag{2.4}$$

$$P(A,B,E,J,M) = P(B) \cdot P(E) \cdot P(A|B,E) \cdot P(J|A) \cdot P(M|A) \tag{2.5}$$

The model in (2.5) fares well on the compactness criterion since it has eight parameters instead of 31. Consider what happens if we choose a different order for our variables, $\langle M,J,A,B,E \rangle$, for example. It turns out that the model in (2.6) is less compact than the one in (2.5), as it is more difficult to simplify it using reasonable assumptions. For example we cannot assume that the probability of John calling is independent of Mary calling (see $P(J|M)$ in (2.6)), since if Mary calls it is likely that the alarm has gone off, which makes it more likely that John calls. We cannot assume that the alarm going off is independent of either John or Mary (see $P(A|M,J)$ in (2.6)), since if both call it is more likely that the alarm has gone off than if just one or neither calls. We can simplify the term $P(B|M,J,A)$, since if we know that the alarm has gone off, a phonecall from Mary or John will not provide further information about the burglary (see (2.7)). We can also simplify the term $P(E|M,J,A,B)$ since knowing that John or Mary called does not contribute much information about the earthquake (see (2.8)). The resulting model is given in (2.9).

$$P(A,B,E,J,M) = P(M) \cdot P(J|M) \cdot P(A|M,J) \cdot P(B|M,J,A) \cdot P(E|M,J,A,B) \tag{2.6}$$

$$P(B|M,J,A) = P(B|A) \tag{2.7}$$

$$P(E|M,J,A,B) = P(E|A,B) \tag{2.8}$$

$$P(A,B,E,J,M) = P(M) \cdot P(J|M) \cdot P(A|M,J) \cdot P(B|A) \cdot P(E|A,B) \tag{2.9}$$

Note that the model in (2.9) has now 12 parameters. However, this new ordering requires the estimation of probabilities for which finding evidence may be problematic. For example, it is not straightforward how to estimate the probability of an earthquake given that a burglary occurred and the alarm went off (see the term $P(E|A,B)$) in (2.9)).

The example illustrates some important lessons for probabilistic models defined as joint distributions: (a) the representation of the problem (i.e., choice of variables) matters (choosing to represent the fact that Mary listens to loud music or that John cannot tell the difference between the phone and the alarm would have resulted in a different and less compact model), (b) the ordering of the variables has an immediate effect on the number of the model parameters and the ease of their estimation, and (c) independence assumptions are constrained by the nature of the problem (it is reasonable to assume that earthquakes are independent from burglaries, but it is not reasonable to assume that John's phone call is independent from Mary's phone call).

Throughout this thesis word combinations and their meanings will be also modeled as joint distributions. Unlike the burglary problem discussed above, our variables will have little to do with real-world events linked via causal relations. We will typically assume variables whose interdependencies arise from linguistic generalizations concerning the syntactic, semantic, and morphological properties of words. More formally, our goal will be to find the most likely meaning M for a word combination W out of the set of allowable meanings $\mu(W)$. In other words, we will assign a probability to the pair (M,W):

$$\hat{M} = \underset{M \in \mu(W)}{\operatorname{argmax}} P(M,W) \tag{2.10}$$

Our first task will be to choose a representation for M and W (i.e., to choose the variables for M and W). For example, M can be either derived from the corpus, e.g., the meaning can be represented by a verb or a preposition, or provided by an external resource such as WordNet or Roget's thesaurus. As far as W is concerned we will assume throughout this thesis that it represents words that are syntactically related rather than combined arbitrarily. Examples of a syntactic relation are a verb and its complements, a noun modified by an adjective or another noun. We may choose to represent the word combination W in terms of the word forms it consists of, their syntactic relation (e.g., subject, object), or their parts of speech. Once

the number and type of variables are chosen we will specify an ordering for them. The latter directly influences the amount of data needed for the estimation of the model parameters.

Consider for example sentence (2.11a). Suppose we wish to derive the most probable meaning for the expression *finish three cigarettes*. In particular, we expect our model to rank the meaning "finished smoking three cigarettes" higher than the meaning "finished eating three cigarettes". Our first decision concerns the representation of M (and $\mu(W)$) and W for (2.11a). One possible choice for M is WordNet and the meanings it provides for *finish*: complete, end up, terminate, coat, eat up. The combination W can be broken down into three events: V for the finishing event, D for the determiner following *finish* and modifying its object, and N for the noun *cigarettes*. This defines the joint distribution $P(D,M,N,V)$. Choosing the ordering $\langle M,D,V,N\rangle$ produces the model shown in (2.12).

(2.11) a. John finished three cigarettes.
b. John finished smoking three cigarettes.

$$P(D,M,N,V) = P(M) \cdot P(D|M) \cdot P(V|M,D) \cdot P(N|M,D,V) \tag{2.12}$$

There are two problems with choosing WordNet as our representation for M. The first is that none of the meanings listed in WordNet captures the fact that *finished three cigarettes* means "finished smoking" rather than "eating three cigarettes". The other problem is related to the parameterization discussed above. Unless we have a corpus annotated with WordNet meanings, we cannot estimate the parameters $P(M)$, $P(D|M)$, $P(V|M,D)$, and $P(N|M,D,V)$. Another approach to the same problem is to ignore WordNet and try to derive the meaning of (2.11a) directly from the corpus. This assumes a different representation for M. We start by noticing that sentence (2.11a) has the same interpretation as (2.11b). The observation is that despite the fact that *finish* takes an NP as its complement (i.e., *three cigarettes*) (2.11a) is interpreted to mean that an event is finished. Instead of relying on WordNet, we will use the corpus to retrieve this event. In this case, our representation for M is simply the set of verbs which can be both complements of *finish* and predicates for *three cigarettes*. Assuming that we choose the ordering $\langle M,D,V,N\rangle$ for our variables, we end up with the same model as in (2.12), except that now it is easier to estimate the model parameters.

As in the case of the alarm-burglary problem, we can further simplify the model in (2.12) using not so much knowledge about causality but knowledge about syntactic and lexical semantic dependencies. We begin by noticing that knowing the underspecified event M will not contribute much information with regard to the words following it. The variable D can be any modifier (e.g., a determiner, an adjective, a quantifier). We conclude thus that D is independent from M and simplify $P(D|M)$ as shown in (2.13). We further observe that the verb V (i.e., *finish*) is independent of the determiner D (e.g., *three*). The meaning of (2.11a) remains the same irrespectively of whether John smoked one, two, or many cigarettes. Following this

reasoning we can simplify the term $P(V|M,V)$ as shown in (2.14). Furthermore, N, the object of *finish* (i.e., *cigarettes*) is more closely related to the underspecified event M rather than to D or V. For example, smoking cigarettes is more likely than eating cigarettes, irrespectively of whether we have finished or begun smoking them or whether the cigarettes are three or expensive. We simplify the term $P(N|D,V,M)$ as shown in (2.15). These simplifications result in a compact model with a relatively small number of parameters (see (2.16)).

(2.13) $P(D|M) = P(D)$

(2.14) $P(V|M,D) = P(V|M)$

(2.15) $P(N|M,D,V) = P(N|M)$

(2.16) $P(D,M,N,V) = P(M) \cdot P(D) \cdot P(V|M) \cdot P(N|M)$

Note that in choosing the variables for (2.11a) we decided not to represent the fact that the finishing event has a subject (i.e., *John*, see (2.11a)). However, we chose to explicitly represent the fact that the object of the verb *finish* may be modified. An alternative representation is to abstract over the particular structure of the object NP by choosing to represent only its head noun. It is straightforward to see that this choice will produce a far less restrictive model which will be able to deliver the most likely meaning for any type of complement NP (irrespectively of whether it is pre- or post-modified). As a result, sentence (2.11a) will be now represented by three variables (i.e., M, N, and V) instead of four (see equation (2.12)). Choice of the ordering $\langle M,V,N \rangle$ produces the model shown in (2.17). As explained previously, we can further simplify the term $P(N|V,M)$ (see (2.18)) by assuming that the object N is independent from the predicate V. The model in (2.19) is more general than the model in (2.16) and has fewer parameters.

(2.17) $P(M,N,V) = P(M) \cdot P(N|M) \cdot P(V|M,N)$

(2.18) $P(V|M,N) = P(V|M)$

(2.19) $P(M,N,V) = P(M) \cdot P(N|M) \cdot P(V|M)$

Note further that the model in (2.19) is ignorant with respect to the discourse context within which sentence (2.11a) is embedded. This means that it will come up with the same

meaning for (2.11a), repeated here as (2.20b) and (2.21b), irrespectively of whether it is preceded by sentence (2.20a) or (2.21a). In other words, the model will predict that a sentence like (2.20b) is likely to mean (2.21c) even in cases where discourse context triggers an alternative interpretation (see (2.20c)). The model thus does not focus on the meaning of individual corpus tokens; instead it determines the most dominant meaning for a given word combination overall, across all of its instances in the corpus.

(2.20) a. Who is making the cigarettes for tomorrow's party?
b. John finished three cigarettes.
c. John finished making three cigarettes.

(2.21) a. Why is the room filled with smoke?
b. John finished three cigarettes.
c. John finished smoking three cigarettes.

The use of joint distributions has been also proposed for supervised word sense disambiguation (see Section 1.2.2 for discussion on word sense disambiguation). Bruce and Wiebe (1999) represent the different senses of a given word as the joint distribution of a variety of features (i.e., part-of-speech tags, morphological features, or collocations). The relationships among the different variables are expressed as a product of marginal distributions, where each marginal distribution is composed by interdependent variables. Depending on the ordering of the variables and the conditional independence assumptions that could be made to simplify the expression of the joint distribution, there is often a variety of models that can represent the joint distribution in question. These so-called *decomposable models* are graphical models, i.e., the interdependencies among the different variables can be expressed graphically in a dependency graph (Whittaker 1990). Given that not one but several probability models represent the phenomenon at hand, a search through the space of probability models is necessary in order to find the model with the fewest interdependencies that fits the data well.

Although we model the dominant meaning of a given word combination also as a joint distribution and use conditional independence assumptions to simplify the model parameters, we assume a particular model rather than searching for one that is appropriate for the data. Our aim is to choose the optimal model on the basis of two important criteria: (a) linguistic faithfulness (i.e., we want our model to capture linguistic interdependencies) and (b) data availability; in contrast to Bruce and Wiebe (1999), our approach is unsupervised and therefore we have to ensure that our model's parameters can be estimated without recourse to manual annotation.

To summarize, in this thesis we will model meaning as a joint distribution of syntactically related linguistic objects. Our task will be to choose the right representation for these objects and to simplify their interdependencies so as to restrict the parameter space. In some cases, we will rely solely on the corpus to provide a representation for our linguistic objects (see Chapter 5). In other cases, we will use a combination of information internal and external to the corpus (see Chapters 4 and 7). The independence assumptions will be crucial for the

performance of our model. In some instances our independence assumptions will be linguistically plausible, in other cases the motivation for these assumptions will be driven by data considerations. In other words, it will be easier to estimate the parameters of a model which crudely represents a linguistic phenomenon than the parameters of a model which completely describes it. In some cases, even with reasonable independence assumptions, we will simply not have enough data to model the phenomenon at hand. We will show that the corpus is a very useful resource for "recreating" the missing evidence (see Chapters 6 and 7).

2.5. Evaluation

Throughout this thesis we will evaluate our results against humans. In research on language and language behavior subjects are often asked to rate or judge linguistic phenomena. Various kinds of judgments can be elicited, and they can be expressed in various ways. In our case, the judgments will relate to the output of an automatic procedure. Subjects will be typically asked to assess the precision of this procedure by categorizing corpus tokens as positives (true, false) or negatives (true, false). In some cases, evaluation of the results of our modeling studies will involve comparison against information not directly available in the corpus. In this case, subjects will be asked to annotate corpus tokens with a set of available meanings against which the results of our model will be compared. We will not only measure the precision of our model against the judges but also the degree to which our judges agree among themselves. Inter-judge agreement will give us an indication of the feasibility of the task. If judges cannot agree on their category assignments, it is unlikely that these can be reproduced automatically. We describe our method for measuring inter-subject agreement in Section 2.5.1. In some cases, evaluation may rely crucially on subtle linguistic intuitions, viz. on judgments of the relative plausibility of different meanings for a given word combination. Such relative acceptability judgments will be measured experimentally. We briefly describe our experimental paradigm for eliciting linguistic judgments in Section 2.5.2.

2.5.1. The Kappa Statistic

The majority of the evaluation studies conducted in this thesis will involve mutually exclusive categorial assignment. We will report percentage agreement, i.e., the percentage of identical categorizations between our judges and an automatic procedure as well as pairwise agreement among our judges. Although percent agreement is an important and intuitive measure of the performance of an algorithm or a probabilistic model, it does not take into account the expected chance inter-rater agreement. Knowing how well the judges agree on their classifications by taking into account the agreement that would be expected if the decisions made by each judge were statistically independent is crucial for our tasks which involve subtle linguistic judgments about meaning and syntax.

The amount of agreement we would expect the judges to reach by chance depends on the number and relative proportions of the categories used by the judges. Consider the following examples taken from Carletta (1996: 250). Assume there are two judges who assign linguistic objects to two categories randomly. If the two categories are equally frequent, the coders would agree with each other half of the time. If the categories were four (instead of two) and equally distributed we would expect them to agree 25% of the time. Consider now what happens if the coders were to use two categories, but one of them 95% of the time. In this case, we would expect them to agree 90.5% of the time (i.e., if the first coder chooses the first category 95% of the time and the second category 5% of the time, there is a .95 chance of the second coder to choose the first category and .05 chance to choose the second category, $.95^2 + .05^2$).

In this thesis we use the Kappa coefficient to measure inter-rater agreement among a set of coders making category judgments. Kappa effectively measures agreement by factoring out expected chance agreement (Cohen 1960). The Kappa coefficient is the ratio of the proportion of times $P(A)$ that k raters agree (corrected by $P(E)$ the proportion of times we would expect them to agree by chance) to the maximum proportion of times the raters would agree (corrected for chance agreement) (see (2.22)). If there is a complete agreement among the raters, then $K = 1$, whereas if there is no agreement among the raters (other than the agreement which would be expected to occur by chance), then $K = 0$ (see Siegel and Castellan 1988 for details on how to calculate K). If two judges agree less than expected by chance, Kappa can be negative.

$$K = \frac{P(A) - P(E)}{1 - P(E)} \tag{2.22}$$

Landis and Koch (1977) give the following five qualifications for different values of Kappa: .00–.20 is slight, .21–.40 is fair, .41–.60 is moderate, .61–.80 is substantial, whereas .81–1.00 is almost perfect.

The use of Kappa in evaluation studies within the field of computational linguistics is becoming increasingly popular. For example, Carletta (1996) and Hirschberg and Nakatani (1996) use the Kappa statistic to measure inter-rater reliability at recognizing discourse boundaries. Wiebe, O'Hara, Öhrström Sandgren, and McKeever (1998) use Kappa to investigate how well coders agree at annotating temporal units, Teufel (2000) uses Kappa to evaluate the performance of a summarization system, Hatzivassiloglou, Klavans, and Eskin (1999) use Kappa to measure inter-rater reliability at judging the similarity of textual units (i.e., paragraphs), and Stevenson and Merlo (2000) use Kappa to assess how well humans agree at classifying verbs into lexical semantic classes.

2.5.2. Magnitude Estimation

Our evaluation studies will not only focus on categorial judgments. In some cases, we will measure the perceived plausibility of a meaning for a certain word combination. Consider

again example (2.11a) from Section 2.4. Assume our model comes up with three potential interpretations for (2.11a): John finished "smoking", "making", or "eating" three cigarettes. In this case, we wish to measure how subjects perceive these meanings as potential interpretations for (2.11a).

An experimental paradigm for eliciting linguistic judgments is magnitude estimation (ME), a technique standardly used in psychophysics to measure judgments of sensory stimuli (Stevens 1975). The magnitude estimation procedure requires subjects to rate the perceived magnitude of physical stimuli by assigning values on an interval scale (e.g., numbers or line lengths) proportional to the magnitude of a modulus item. Highly reliable judgments can be achieved in this fashion for a wide range of modalities, such as brightness, loudness or tactile stimulation.

The ME paradigm has been applied by Bard, Robertson, and Sorace (1996), Cowart (1997) and Keller (2000) to the elicitation of linguistic judgments. ME has been shown to provide fine-grained measurements of linguistic acceptability which are robust enough to yield statistically significant results, while being highly replicable both within and across speakers. ME has been applied to several linguistic phenomena such as auxiliary selection, coordination and binding, resumptive pronouns, extraction, unaccusativity, gapping, word order, and selectional restrictions (see Keller 2000 for an overview).

The ME procedure for linguistic acceptability is analogous to the standard procedure applied to the elicitation of judgments for physical stimuli. Subjects are presented with a series of linguistic stimuli, and have to respond by assigning a value to each stimulus proportional to the acceptability they perceive. Subjects are first exposed to a modulus item, to which they assign an arbitrary number. All other stimuli are rated proportional to the modulus. In this way, each subject can establish their own rating scale, thus yielding maximally fine-grained data and avoiding the known problems with the conventional ordinal scales for linguistic data (Bard et al. 1996; Cowart 1997; Schütze 1996).

We used the ME paradigm in Chapter 5 to evaluate our model's performance at deriving meanings for polysemous adjective-noun combinations by comparing the model's rankings against judgments of meaning paraphrases elicited experimentally from human subjects.

2.6. Summary

In this chapter we laid the methodological foundations for the acquisition and modeling studies carried out in the remainder of the thesis. We presented the corpora used in our studies and the parsing methodology (partial parsing and regular expression matching) employed for the extraction of syntactic information. We also gave an overview of the lexical resources (taxonomies, morphological, and subcategorization dictionaries) used throughout the thesis. We presented the probabilistic model underlying our modeling studies and discussed its properties

(choice of model variables and their ordering, independence assumptions, parameter estimation, and model compactness). We also described our methodology (percentage agreement, inter-judge agreement, Magnitude Estimation) for evaluating our modeling and acquisition results.

Chapter 3

Diathesis Alternations

This chapter presents our first acquisition study which focuses on verbs and their subcategorization. We attempt to validate the claim that lexicon acquisition and linguistic theory can be complementary by starting from Levin's (1993) theory of diathesis alternations and examining the extent to which these are attested in corpus data. We acquire frames characteristic of diathesis alternations from the British National Corpus (BNC) using the *surface cueing* approach and show how the acquired subcategorization type and token frequencies can be used to empirically measure whether an alternation is *productive* (i.e., whether the verbs listed in Levin are found to alternate in the corpus) and whether a verb or verb class are *typical* (i.e., representative) of the alternation. We examine the validity of the typicality and productivity measures as formalizations of the empirical behavior of alternating verbs by testing their predictions in the Penn Treebank corpus. The results based on the latter corpus largely confirm the typicality and productivity predictions obtained from the BNC, suggesting the usefulness of these two empirical measures. Finally, the experiments in this chapter demonstrate that our approach is not only useful for validating Levin's claims about the behavior of alternating verbs but also for discovering novel verbs licensing a given alternation.

3.1. Introduction

Diathesis alternations are changes in the realization of the argument structure of a verb that are sometimes accompanied by changes in meaning (Levin 1993). The phenomenon in English is illustrated in (3.1)–(3.6), taken from Levin (1993).

(3.1) a. Janet broke the cup.
b. The cup broke. (Levin 1993: 29)

(3.2) a. Jack sprayed paint on the wall.
b. Jack sprayed the wall with paint. (Levin 1993: 51)

(3.3) a. I filled the pale with water.

b. Water filled the pale. (Levin 1993: 81)

(3.4) a. The boy opened the window.

b. The window just opened itself. (Levin 1993: 84)

(3.5) a. The cook sliced the mushrooms.

b. The mushrooms were sliced by the cook. (Levin 1993: 86)

(3.6) a. A problem developed.

b. There developed a problem. (Levin 1993: 89)

Example (3.1) is an illustration of the causative/inchoative alternation. Verbs undergoing this alternation can be manifested either as transitive with a causative reading (see sentence (3.1a)) or as intransitive with an inchoative reading (see sentence (3.1b)). The Spray/Load alternation is exemplified in (3.2). This alternation is usually licensed by verbs involving two arguments, the locatum argument, i.e., the entity whose location is changed (e.g., *paint* in (3.2a)), and the location argument (e.g., *wall* in (3.2a)). The verb alternates between a variant where the location argument is the object of a prepositional phrase and the locatum argument is the object of the verb (see (3.2a)) and a variant where the location argument is the object of the verb and the locatum argument is the object of the prepositional phrase (see (3.2b)). For verbs that license the Spray/Load alternation the changes in their argument structure are accompanied by meaning changes. The variant where the location argument is the object of the verb receives a "holistic" interpretation which implies that the whole wall was sprayed with paint (see (3.2b)), whereas the variant where the location argument is the object of the preposition receives a "partitive" interpretation which does not imply that the whole wall was sprayed with paint (see (3.2a)).

Example (3.3) illustrates the locatum subject alternation: the argument of the verb can be either expressed with a prepositional phrase (see (3.3a)) or as an oblique subject (see (3.3b)). The reflexive diathesis alternation is exemplified in (3.4): verbs undergoing this alternation can have a typical transitive use (see (3.4a)) or a use where the verb takes a reflexive object (see (3.4b)). Examples (3.5) and (3.6) illustrate the passive and *there*-insertion alternation, respectively. In the *there*-insertion alternation the verb is usually intransitive and alternates between a typical use (see (3.6a)) and a use where *there* appears in the canonical subject position (i.e., before the verb) and the subject appears postverbally (see (3.6b)).

The main assumption behind Levin's (1993) study of diathesis alternations is that the syntactic realization of a verb's arguments is to a large extent determined by its meaning (see also Pinker 1989 for a similar proposal). Hence one would expect that verbs that participate in the same diathesis alternations share certain meaning components and consequently form a semantically coherent class. Levin defines six main types of diathesis alternations for English: (a) transitivity alternations (see (3.1)), (b) alternations involving arguments within the VP (see (3.2)), (c) oblique subject alternations (see (3.3)), (d) reflexive diathesis alternations (see (3.4)), (e) alternations involving passives (see (3.5)), and (f) alternations involving postver-

bal subjects (see (3.6)). Each of these alternation types exhibits a wide variety of sub-types resulting to a total of 79 alternations. Furthermore, Levin defines approximately 200 verb semantic classes, under the assumption that members of a given class pattern together with respect to diathesis alternations.

The objective of this chapter is to examine the extent to which diathesis alternations are attested in corpus data. More specifically, we attempt to determine: (a) if frame alternants can be acquired automatically from corpora using a surface cueing approach, (b) if some alternations are more frequent than others, (c) if alternating verbs have frame preferences, and (d) what the representative members of an alternation are. We focus on surface syntactic structure (i.e., subcategorization) rather than the meaning components of verbs, primarily because syntactic information, in contrast to meaning, can be easily identified and extracted from corpora. We employ a methodology inverse to Levin's (1993): instead of looking for semantic regularities as indicators of common syntactic behavior, we treat the alternative argument structures manifested by a verb as clues about its meaning.

We present the results of a set of experiments which investigate three alternations which involve arguments within the VP (i.e., the dative, benefactive, and possessor object alternation, see the examples in (3.7)–(3.9)) and one transitivity alternation (i.e., the conative alternation, see (3.10)). We discover verbs participating in these alternations automatically by recognizing shallow linguistic patterns, such as basic syntactic relationships between words, in a non-parsed corpus, annotated with part-of-speech information. To achieve approximately correct frame frequencies we further make use of linguistic heuristics and the WordNet taxonomy (Miller et al. 1990). The analysis presented here is based on the British National Corpus (BNC), a 100 million word collection of samples of written and spoken language from a wide range of sources designed to represent a wide cross-section of current British English (Burnard 1995, see Chapter 2 for details).

(3.7) a. Bill sold Tom a car.
b. Bill sold a car to Tom. (Levin 1993: 46)

(3.8) a. Martha carved the baby a toy.
b. Martha carved a toy for the baby. (Levin 1993: 49)

(3.9) a. I admired his honesty.
b. I admired him for his honesty. (Levin 1993: 192)

(3.10) a. Paula hit the fence.
b. Paula hit at the fence. (Levin 1993: 41)

Our study comprises four experiments: Experiment 1 (see Section 3.2) focuses on the dative and benefactive alternation (see (3.7) and (3.8), respectively). The semantic and syntactic properties of these alternations have been extensively studied and are well understood (Boguraev and Briscoe 1989; Briscoe and Copestake 1999; Goldberg 1995; Levin 1993; Pinker

1989). Experiments 2 and 3 (see Sections 3.3 and 3.4, respectively) investigate the less well-known possessor object and conative alternations (see examples (3.9) and (3.10)). All four alternations seem fairly productive, i.e., a large number of verbs undergo these alternations according to Levin. We investigate whether they are also well represented in a large corpus. Finally, Experiment 4 attempts to validate the results of Experiments 1–3 for a different corpus, the Penn Treebank (Marcus, Grace, Marcinkiewicz, MacIntyre, Bies, Ferguson, Katz, and Schasberger 1994, see Chapter 2 for details).

3.2. Experiment 1: The Dative and Benefactive Alternations

3.2.1. Introduction

Verbs undergoing the dative and benefactive alternations pattern with Spray/Load verbs (see the examples in (3.2)) in that they allow more than one way of expressing their arguments. However, they differ from Spray/Load verbs in one important aspect: the changes in the realization of their argument structure are not accompanied by changes in meaning. The dative alternation illustrated in example[1] (3.7) is characterized by an alternation between the double object frame 'NP1 V NP3 NP2' (see sentence (3.7a)) and the prepositional frame 'NP1 V NP2 *to* NP3' (see sentence (3.7b)). The benefactive alternation is structurally similar to the dative alternation, the difference being that it involves the preposition *for* rather than *to* in the prepositional variant (see example (3.8b)). Dative and benefactive constructions display related semantic properties: both constructions involve a volitional agent (e.g., the NPs *Bill* in (3.7) and *Martha* in (3.8)), a recipient (e.g., the NPs *Tom* in (3.7) and *the baby* in (3.8)) and a theme (e.g., the NPs *a car* in (3.7) and *a toy* in (3.8)); in the dative construction the recipient receives the object denoted by the theme whereas in the benefactive construction the recipient may or may not receive it (see (3.8)).

Experiment 1 is a proof of concept and as such focuses on alternations whose syntactic patterns can be easily detected by shallow parsing techniques that allow the efficient processing of large amounts of corpus data. Compare examples (3.7) and 3.11: a purely syntactic parser can distinguish that sentences (3.7a) and (3.7b) are instances of two distinct frames, whereas this is not the case for the examples in (3.11). Sentence (3.11a) is an instance of the fulfilling alternation, (3.11b) is an instance of the locative alternation, and (3.11c) exemplifies the instrument subject alternation.

(3.11) a. John presented the student with an award.
b. John loaded the truck with bricks.
c. John hit the wall with a hammer. (Dorr 1997: 291)

[1] Unless stated otherwise the example sentences were taken from the BNC and simplified for purposes of clarity. Sentences with asterisks are provided by the author.

Table 3.1: Sample of verbs with frames characteristic for the dative and benefactive alternation

Dative	Benefactive	To-only	Double object	For-only
advance	accept	address	ask	build
allocate	acquire	announce	assume	buy
ask	choose	babble	bear	call
assign	collect	cackle	believe	catch
give	create	demonstrate	charge	choose
grant	indicate	elucidate	consider	cut
flip	obtain	introduce	find	design
float	pick	present	make	develop
loan	prefer	provide	prove	draw
kick	produce	refer	refuse	find

Verbs are classified by Levin (1993) as participating in the alternation (i.e., as having both frames[2]) or non-participating (i.e., as having only one frame). For instance, both frames are available for *preach* and *spin* which participate in the dative and benefactive alternation, respectively (see examples (3.12) and (3.13)). In contrast, the verb *call* (see example (3.14)), only subcategorizes for a double object (i.e., there is no prepositional variant), whereas the verb *receive* only takes a prepositional complement (see example (3.15)). Levin lists 115 verbs which undergo the dative alternation and 103 which license the benefactive alternation. A sample of alternating and non-alternating verbs taken from Levin is given in Table 3.1. Under the columns labeled 'Dative' and 'Benefactive' verbs are listed which Levin classifies as participating in the respective alternations. Under the remaining columns non-alternating verbs are listed (i.e., verbs which Levin classifies as having only one of the two alternating frames).

(3.12) a. The priest preached the gospel to them.
b. The priest preached them the gospel.

(3.13) a. John spinned a romantic tale for his girlfriend.
b. John spinned his girlfriend a romantic tale.

(3.14) a. He called me a trouble-maker.
b. *He called a trouble-maker to me.

(3.15) a. Her father received substantial support for his research.
b. *Her father received his research substantial support.

In the following we describe and evaluate the set of automatic methods we used to discover verbs undergoing the dative and benefactive alternations (see Section 3.2.2). We assess the acquired frames using a filtering method presented in Section 3.2.2.7. The results are detailed in Section 3.2.3. Sections 3.2.3.1 and 3.2.3.2 discuss how the derived type and token

[2]In the following we use the term frame to refer to the arguments of verbs undergoing both the dative and benefactive alternation, even though benefactive NPs and PPs are not obligatory (see (3.8)).

frequencies can be used to estimate how productive an alternation is for a given verb semantic class and how typical its members are.

3.2.2. Method

3.2.2.1. Acquisition

A part-of-speech tagged and lemmatized version of the BNC (90M words written and 10M words spoken language) was used for the extraction of the syntactic structures characteristic for the dative and benefactive alternations. Surface syntactic structure was automatically extracted from the BNC using Gsearch (Corley et al. 2001, see Chapter 2 for details). We used Gsearch to extract tokens matching the syntactic patterns in (3.16) by specifying a chunk grammar for the verbal complex and NPs. Part-of-speech tags were retained in the parser's output which was post-processed to remove adverbials and interjections.

(3.16) a. V NP1 NP2
b. V NP1 *to* NP2
c. V NP1 *for* NP2

Examples of the parser's output are illustrated (3.17)–(3.19), where the brackets show the parse Gsearch came up with. Although there are cases where Gsearch produces the right parse (see examples (3.17a,b), (3.18a), and (3.19a)), the parser wrongly identifies as instances of the double object frame tokens containing compounds (see example (3.17c) where the compound *the death penalty* is parsed as two separate NPs, *the death* and *penalty*), bare relative clauses (see example (3.17d)), or NPs in apposition (see example (3.17e)). Sometimes the parser attaches prepositional phrases at wrong sites. This is shown in sentence (3.18b), where the PP *to the vault* modifies the verb *control* instead of being attached to the NP *access*. Finally, even if the parser produces the right parse, it cannot distinguish between arguments and adjuncts (see (3.19b) and (3.19a), respectively) or between different types of adjuncts (e.g., temporal (see (3.19c)) versus benefactive (see (3.19a))).

(3.17) a. I have also campaigned for the Government to [$_{VP}$ give] [$_{NP}$ AIDS] [$_{NP}$ greater recognition], not as a disease affecting specific sectors of the community, but as a social problem for which there must be adequate welfare provision.
b. The police driver [$_{V}$ shot] [$_{NP}$ Jamie] [$_{NP}$ a look of enquiry] which he missed, occupied as he was with guiding Miss Williams back up the hill.
c. The definition of capital murder which [$_{V}$ carries] [$_{NP}$ the death] [$_{NP}$ penalty] varies from state to state.
d. A Jaffna schoolboy [$_{V}$ shows] [$_{NP}$ a drawing] [$_{NP}$ he] made of helicopters strafing his home town.

e. For the latter catalogue Barr [$_V$ chose] [$_{NP}$ the Surrealist writer] [$_{NP}$ Georges Hugnet] to write a historical essay, but it was Barr's intelligent advocacy and choice of works that made the show such a significant affair.

(3.18) a. However, before competing again you must [$_V$ present] [$_{NP}$ yourself] [$_{PP}$ to the tournament doctor] to have both the injury and your bandaging accepted.

b. It was clear that the premises were used for viewing works of art stored there by Capricorn and that it [$_V$ controlled] [$_{NP}$ access] [$_{PP}$ to the vault].

(3.19) a. I said to David, I think you'll have to have your hair cut, otherwise you won't get this programme on television, to which he replied that he wouldn't [$_V$ cut] [$_{NP}$ his hair] [$_{PP}$ for the Prime Minister].

b. The attendant [$_V$ mistook] [$_{NP}$ him] [$_{PP}$ for his rival].

c. Yesterday he [$_V$ rang] [$_{NP}$ the bell] [$_{PP}$ for a long time].

(3.20) a. Kay, an ACET nurse, visits a client to administer Petamadine to help [$_V$ prevent] [$_{NP}$ him] [$_{NP}$ developing pneumocystis corinii pneumonia] (PCP).

b. A new brand of screen sponsorship: Easing of broadcasting rules will [$_V$ let] [$_{NP}$ advertisers] [$_{NP}$ fund TV programmes].

Erroneous output is also due to tagging mistakes. The parser produced sentence (3.20a) as a match to the syntactic pattern 'V NP NP': since *developing* was wrongly tagged as an adjective instead of a gerund, the parser incorrectly classified the sequence *developing pneumocystis corinii pneumonia* as a noun phrase. The tag NN1-VVB (NN1 stands for singular common noun and VVB for the finite base form of verbs other than auxiliaries) was assigned to the verb *fund* in (3.20b). As a result, the parser recognized the sequence *fund TV programmes* as a noun phrase.

We treated tagging errors as noise which we do not have the means to eliminate. Given that some parses are incorrect, we cannot distinguish whether patterns of the form 'V NP1 NP2' are instances of a verb which takes two arguments, namely the two noun phrases, or whether such patterns are instances of a verb which takes one argument, namely the two noun phrases constitute a single noun phrase as in the case of compounds or proper names. Similarly, given parsing errors, we cannot be certain of the attachment site of the prepositional phrase in patterns of the form 'V NP1 *to* NP2' or 'V NP1 *for* NP2' (the prepositional phrase can be attached to the verb or the noun following it). We identified erroneous subcategorization frames (see (3.17c)–(3.17e)) by using linguistic heuristics and a process for compound noun detection. We disambiguated the attachment site of PPs (see (3.18b)) using Hindle and Rooth's (1993) lexical association score. Finally, we recognized benefactive PPs (see (3.19a)) by exploiting the WordNet taxonomy (see Chapter 2 for details).

3.2.2.2. Guessing the double object frame

We developed a process which assesses whether the syntactic patterns (called cues below) derived from the corpus were instances of the double object frame. The method is a combination of linguistically motivated heuristics and corpus-based statistics.

Linguistic heuristics. We applied several heuristics to the parser's output which determined whether corpus tokens were instances of the double object frame. The 'Reject' heuristics below identified erroneous matches (i.e., false positives, see (3.17c)–(3.17e)), whereas the 'Accept' heuristics identified true instances of the double object frame (i.e., true positives, see (3.17a,b), (3.18a), and (3.19a))).

1. **Reject** if cue contains at least two proper names[3] adjacent to each other (e.g., *killed Henry Phipps*, *defend Saudi Arabia*).

2. **Reject** if cue contains possessive[4] noun phrases (e.g., *affect the body's defence system*, *ask God's pardon*, *give a showman's award*).

3. **Reject** if cue's last word is a personal or reflexive pronoun (e.g., *liked the way he*, *tells Carol she*, *ask the subjects themselves*).

4. **Accept** if verb is followed by a personal or indefinite pronoun (e.g., *gave them a three-hour lecture*, *allows you another opportunity*, *found him a new home*, *making everyone a shareholder*, *buy someone a drink*).

5. **Accept** if verb is followed by a reflexive pronoun (e.g., *made herself a snack*, *give myself injections*, *owe each other duties*).

6. **Accept** if cue's surface structure is 'V MOD* NP MOD[5] NP' (e.g., *send Bailey a postcard*).

7. **Cannot decide** if cue's surface structure is 'V MOD* N N+' (e.g., *offer a free bus service*).

Compound Nouns. Examples of tokens identified by heuristic 7 are given in Table 3.2. These tokens were dealt with separately by a procedure which guesses whether the nouns following the verb are two distinct arguments or parts of a compound (see (3.17c)). This procedure was applied only to noun sequences of length two and three. Tokens containing noun sequences of length larger than three (450 in total) were rejected as instances of the double object frame.

[3]Proper nouns are annotated in the BNC corpus distinctly from common nouns and therefore can be easily identified.

[4]The possessive marker *s* is annotated with a separate part-of-speech tag in the BNC and therefore possessive NPs can be easily identified.

[5]Here MOD represents any prenominal modifier (e.g., articles, pronouns, adjectives, quantifiers, ordinals).

Table 3.2: A sample of noun sequences and their preceding verbs extracted from the parsed corpus

Cues	Examples
V N N	granted market monopolies
V MOD ADJ N N	offer a free bus service
V MOD N N	shows some example words
V N N N	give service duties priority
V MOD ADJ N N N	bring their respective application development tools
V MOD N N N	want some envelope re-use labels

From the tokens identified by heuristic 7 we extracted noun sequences of length two (196,464 in total) and three (23,540 in total). These noun sequences were compared against a dictionary of compound nouns which was compiled from WordNet (Miller and Charles 1991) and contained 48,661 entries. 13.9% of the noun sequences attested in our data were found in WordNet (25,251 sequences of length two and 5,389 noun sequences of length three). Tokens containing these sequences were eliminated on the basis of the assumption that the nouns following the verbs in question were established compounds and could not therefore be verbal arguments.

For the remaining 171,213 sequences of length two we used the log-likelihood ratio (LLRatio) to estimate the lexical association between the nouns, in order to determine if they formed a compound noun. We assumed that two nouns cannot be disjoint arguments of a verb if they are lexically associated. On the basis of this assumption cues were rejected (i.e., not classified as instances of the double object frame) if they contained two nouns whose log-likelihood ratio had a p-value less than .05. We preferred the log-likelihood statistic (G-score) to other statistical scores, such as the association ratio (Church and Hanks 1990) or χ^2, since it adequately takes into account the frequency of the co-occurring words and is less sensitive to rare events and corpus-size (Daille 1996; Dunning 1993).

We detected noun sequences of length three using a similar, yet more complex, process. We assumed that a verb followed by three nouns either subcategorizes for two arguments, one of which is a compound, or that it takes one argument which is a three word compound noun. Consider for example the cue *bring their respective application development tools* in Table 3.2. Here *application development tools* can be analyzed as [$application_N$ [$development_N$ $tools_N$]] or [[$application_N$ $development_N$] $tools_N$]. Under this assumption, we first determined the bracketing of three word noun sequences and then computed the log-likelihood ratio between a single noun and a two word noun sequence.

We inferred the bracketing by modifying an algorithm initially proposed by Pustejovsky et al. (1993). The algorithm assumes a three word compound as its input and searches the corpus for its possible subcomponents. Whichever subcomponent is found is chosen as the more closely bracketed pair. Pustejovsky et al. do not explain how the algorithm behaves when

Given three nouns n_1, n_2, n_3:

1. **if** either $[n_1\ n_2]$ or $[n_2\ n_3]$ appear in WordNet **then**
   build structures $[[n_1\ n_2]\ n_3]$ or $[n_1\ [n_2\ n_3]]$ accordingly;
2. **if** both $[n_1\ n_2]$ and $[n_2\ n_3]$ appear in WordNet **then**
   choose most frequent pair;
3. **if** neither $[n_1\ n_2]$ nor $[n_2\ n_3]$ appear in WordNet **then**
   (a) compute LLRatio for $[n_1\ n_2]$ and $[n_2\ n_3]$;
   (b) **if** $p \leq .05$ for both pairs **then**
   choose pair with highest LLRatio;
   (c) **if** $p \geq .05$ for both pairs **then**
   choose pair with highest relative frequency;
   (d) **if** both pairs have equal relative frequencies **then**
   cannot decide bracketing.

Figure 3.1: Algorithm for bracketing noun sequences of length three

neither or both subcomponents appear in the corpus. Furthermore, bracketing is determined simply by appearance of a two word noun sequence in the corpus without taking its corpus frequency into account.

In Figure 3.1 we present our modified version of the algorithm. The input to the algorithm, is not an already established three word compound as in Pustejovsky et al. (1993), but any noun sequence of three words occurring after a verb. The algorithm determines the internal structure of the candidate noun sequence by combining the relative corpus frequency of its sub-parts (see steps 1–3 in Figure 3.1) and information about established compounds (i.e., WordNet). Once the internal structure of the three word sequence is determined[6], the log-likelihood ratio is used so as to evaluate whether it is a compound or not. The log-likelihood ratio is computed between a noun and a noun sequence of length two. As in the case of noun sequences of length two, cues are not classified as instances of the double object frame if they contain nouns whose log-likelihood ratio has a p-value less than .05.

Consider again the sequence *application development tools* from Table 3.2. Neither *application development* nor *development tools* are listed as compounds in the dictionary compiled from WordNet. The log-likelihood ratio for the sequence *application development* is higher than for *development tools* (716.49 ($p < .01$) and 371.85 ($p < .01$), respectively). As a result the algorithm chooses the bracketing [[application$_N$ development$_N$] tools$_N$] (see steps (3a,b) in Figure 3.1). The log-likelihood ratio for *application development* and *tools*

[6]The bracketing could not be determined for 918 tokens (see step (3d) in Figure 3.1). These were eliminated from further computations.

Table 3.3: Random sample of two word compounds discovered by the log-likelihood ratio

LLRatio	Compound
1967.7	bank manager
775.2	tax liability
87.0	income tax
75.9	community rule
45.4	book reviewer
30.6	designer gear
29.9	safety plan
24.0	drama school
16.6	airline stewardess
16.5	sales resistance

Table 3.4: Random sample of three word compounds discovered by the log-likelihood ratio

LLRatio	Compound
574.5	[[energy efficiency] office]
382.9	[[council tax] bill]
77.8	[alcohol [education course]]
48.8	[hospital [out-patient department]
36.4	[[tumour suppressor] function]
32.4	[[nature conservation] resource]
24.0	[[quality amplifier] circuit]
22.5	[[language acquisition] device]
20.5	[[Sunday afternoon] concert]
14.7	[ozone [pollution incident]]

is 217.63 ($p < .01$) and consequently the cue *bring their respective application development tools* is rejected as an instance of the double object frame. Tables 3.3 and 3.4 display a random sample of the compounds the method found ($p \leq .05$).

3.2.2.3. Evaluation

The performance of the linguistic heuristics and the compound noun detection procedure were evaluated by randomly selecting approximately 3,000 corpus tokens which were previously accepted or rejected as instances of the double object frame. Two judges were asked to decide whether the corpus tokens were instances of the double object frame. In cases where classification was ambiguous or problematic they were instructed to use the full sentence context. The judges were given minimal instructions (basically examples of verbs with double objects) and no prior training. The judges' agreement on the classification task was measured using the Kappa coefficient (Cohen 1960) which measures inter-rater agreement among a set of coders making category judgments (see Chapter 2 for details).

Table 3.5 reports the precision of the heuristics which classify a given token as an instance of the double object frame (Accept) as well as the precision of the heuristics which reject tokens as instances of the double object frame (Reject). We report average precision (by computing the mean of the precision of both judges) and inter-judge agreement. The 'Reject' heuristics classified correctly 96.6% of the sample tokens, whereas the 'Accept' heuristics achieved an accuracy of 73.6%. The judges reached $K = .76$ ($N = 1,000$, $k = 2$) on the 'Reject' heuristics and $K = .82$ ($N = 1,000$, $k = 2$) on the 'Accept' heuristics.

In sum, both the 'Accept' and 'Reject' heuristics achieved a high accuracy in classifying cues for the double object frame. The accuracy for the 'Accept' heuristics is lower, however.

Table 3.5: Precision of the heuristics for the double object frame and the compound detection procedure

	Precision	K
Accept V NP NP	73.6%	.82
Reject V NP NP	96.6%	.76
2 word compounds	98.9%	.70
3 word compounds	99.1%	

This is mainly due to 'Accept' heuristic 6 (i.e., Accept if cue's surface structure is 'V MOD* NP MOD NP') which overgenerates: of the 247 tokens classified as false positives by the first judge, 231 were due to heuristic 6; similarly, due to same heuristic 255 tokens were classified as false positives by the second judge. Agreement on the classification was good considering that the judges were given minimal instructions and no prior training.

We further evaluated the compound detection procedure by randomly selecting 1,000 tokens (containing approximately 500 noun sequences of length two and 500 noun sequences of length three) which the method rejected as instances of the double object frame. These were manually inspected and classified (as compounds or non-compounds) by two judges. Precision figures (i.e., the mean of the precision of both judges) are reported in Table 3.5. The method performed well in recognizing tokens which were not instances of the double object frame reaching an accuracy of 98.9% for two word noun sequences and 99.1% for three word noun sequences. The judges' agreement on the classification task was $K = .70$ ($N = 998$, $k = 2$).

3.2.2.4. The prepositional frames

In order to consider verbs in prepositional frames as candidates for the dative and benefactive alternations the following requirements need to be met:

1. the prepositional phrase must be attached to the verb;
2. in the case of the 'V NP1 *to* NP2' structure, the *to*-PP must be an argument of the verb;
3. in the case of the 'V NP1 *for* NP2' structure, the *for*-PP must be benefactive;

Syntactically speaking, *for*-PPs are not arguments but adjuncts. Benefactive *for*-PPs are optional (Jackendoff 1990; Wechsler 1995) and can appear on any verb with which they are semantically compatible (Pollard and Sag 1987). In contrast, *to*-PPs are not optional but required by the semantics of the verbs in question. Throughout this chapter we use the term frame to also refer to prepositional structures with benefactive PPs, even though the verb does not strictly subcategorize for a *for*-PP. In order to meet requirements 1–3, we first determined

Table 3.6: A sample of verbs and head nouns followed by PPs

	Verb	Noun	Preposition	Noun
a.	accompany	Castro	to	States
b.	buy	product	for	child
c.	make	trip	to	Germany
d.	experience	preference	for	silence
e.	ask	her	for	advice
f.	say	anything	to	Susan
g.	invite	each other	to	meal
h.	open	door	for	her
i.	give	lift	to	anyone
j.	build	security	for	themselves
k.	bring	baby	for	baptism
l.	dominate	industry	for	year
m.	sue	referee	for	libel

the attachment site (i.e., verb or noun) of the prepositional phrase and secondly developed a procedure for distinguishing benefactive from non-benefactive PPs.

From the syntactic analysis provided by the parser we extracted a table containing the verb and the head of the noun phrase following it (see Table 3.6). For each noun phrase head we recorded the following prepositional phrase ignoring whether or not the parser had attached the preposition to the noun phrase or the verb. Only the head of noun phrases was entered in the table. In the case of compounds we considered as head the rightmost occurring noun (Spencer 1991). Entries in Table 3.6 do not indicate whether the prepositional phrase is attached to the verb or the noun following it. Cases (c) and (d) are examples of a prepositional phrase modifying a noun phrase, whereas cases (a)–(b) and (e)–(m) are examples of verb attachment. Note that verb attachment does not necessarily mean that the prepositional phrase is benefactive or an argument of the verb (see cases (k)–(m) in Table 3.6).

Several approaches have statistically addressed the problem of prepositional phrase ambiguity, with comparable results (Brill and Resnik 1994; Collins and Brooks 1995; Hindle and Rooth 1993; Ratnaparkhi 1998). Hindle and Rooth (1993) used a partial parser to extract $\langle v, n, p \rangle$ tuples from a corpus, where p is the preposition whose attachment is ambiguous between a verb v and a noun n. Hindle and Rooth proposed that ambiguous prepositional phrase attachments can be resolved on the basis of relative strength of association with verbal and nominal heads, estimated on the basis of distribution in an automatically parsed corpus.

Their algorithm starts by exploiting cases where there is no ambiguity with respect to the attachment of the PP. For example, PPs following pronouns which in turn are preceded by verbs (see cases (e)–(g) in Table 3.6) are treated as unambiguous cases of verb attachment. Similarly, PPs which follow sentence initial nouns are considered as evidence of noun attachment (see example (3.21a)). These provide an initial estimate of how likely it is for a preposition to

attach to a verb or noun. For the ambiguous cases, a log-likelihood ratio, the lexical association score (LA-score, see equation (3.22)), compares the attachment probabilities using evidence from the unambiguous cases and decides the attachment site: tuples with an LA-score greater than 2 are assigned to verb attachment, whereas tuples with LA-score less than -2 are assigned to noun attachment. The remaining ambiguous tuples, are classified both as noun and verb attachments by assigning respectively a count of .5 to the noun-preposition and verb-preposition pair.

(3.21) a. Application for a grant should be made at the same time as the application for an audition.

b. A member of a licensing board supported his application for a grant.

Hindle and Rooth's (1993) approach crucially relies on information pertaining to the structural position of PPs in the corpus. Assume for instance that we want to decide the attachment site of the PP *for a grant* in sentence (3.21b). The fact that the same PP has been previously seen in a subject position, i.e., attached to a subject noun (see (3.21a)), favors a noun attachment decision. Such data about the structural position of PPs in the corpus was not available to us, except in cases where the PP followed the verb (see Table 3.6). The present experiment narrowly focused on corpus tokens with surface syntactic structure resembling the frames characteristic for the dative and benefactive alternation (see Section 3.2.2.1). As a result, we could not strictly replicate the procedure detailed in Hindle and Rooth: we had no estimates about the occurrences of PPs preverbally and hence we could produce no initial estimates as to whether a given PP attaches to the noun or not. Consequently we used a variant of the method described in Hindle and Rooth which makes decisions about the attachment site of PPs in an unsupervised, non-iterative manner. Furthermore, the procedure was applied to the special case of tuples containing the prepositions *for* and *to* only.

Following Hindle and Rooth (1993) we assumed that there is a forced choice between two outcomes: the preposition attaches either to the verb or the noun. First we considered the cases where there is no ambiguity with respect to the attachment site for the preposition. Heuristics 1 and 2 separated these cases from the rest of the data.

1. **Verb Attach** if the noun phrase head following the verb is a personal, indefinite, or reflexive pronoun (see cases (e)–(g) in Table 3.6).

2. **Verb Attach** if the noun phrase head following the preposition is a personal, indefinite or reflexive pronoun (see cases (h)–(j) in Table 3.6).

For the remaining data, we used Hindle and Rooth's (1993) lexical association score to compare the attachment probabilities by looking at the probability of the ratio of the preposition attaching to the verb to the probability of the preposition attaching to the noun (see equation (3.22)). Equations (3.23) and (3.24) estimate the probability that the preposition is attached to verb and

noun, respectively. The term $P(NULL|n)$ in (3.23) captures the fact that a PP can be attached to the verb only if is not attached to the noun, i.e., only if the noun following the verb has a low probability of being followed by a preposition in the corpus. There is no such requirement for the PP when it is attached to the noun (see equation (3.24)). The estimation of the quantities $P(p|v)$, $P(NULL|n)$, and $P(p|n)$ is shown in (3.25)–(3.28).

$$LA(v,n,p) = \log_2 \frac{P(verb\_attach\, p|v,n)}{P(noun\_attach\, p|v,n)} \tag{3.22}$$

$$P(verb\_attach p|v,n) \approx P(p|v) \cdot P(NULL|n) \tag{3.23}$$

$$P(noun\_attach p|v,n) \approx P(p|n) \tag{3.24}$$

$$P(p|v) = \frac{f(v,p)}{f(v)} \tag{3.25}$$

$$P(p|n) = \frac{f(n,p)}{f(n)} \tag{3.26}$$

$$P(NULL|n) = \frac{f(n,NULL)}{f(n)} \tag{3.27}$$

$$f(n,NULL) \approx f(n) - f(n,P) \tag{3.28}$$

Table 3.6 was used to derive bigram counts of verb and preposition pairs (see (3.25)). However, Table 3.6 is only partially representative of the frequency with which a given noun occurs with an immediately following preposition—here only the noun-preposition pairs which follow the verbs are attested. Bigram counts for noun-preposition pairs were collected from the entire BNC using Gsearch (Corley et al. 2001). Recall that since we did not parse the entire corpus we did not have a direct estimate of the number of times a given noun does not co-occur with a preposition (see (3.27)). We indirectly estimated the term $f(n,NULL)$ by subtracting from the word's unigram count the count of the pair consisting of the noun and the prepositions following it, represented by $f(n,P) = \sum_p f(n,p)$ (see 3.28). Small frequencies were adjusted as described in Hindle and Rooth (1993) by using Laplace's M estimate. The sign of the LA score was used to indicate attachment preferences: a positive LA score means verb attachment, whereas a negative score means noun attachment.

Table 3.7: Precision of the lexical association procedure and syntactic function of *to*-PPs

	Precision	*K*
Verb Attach-to	74.4%	.78
Noun Attach-to	80.0%	.80
Argument-to	73.4%	.90

3.2.2.5. Evaluation

We evaluated the procedure by randomly selecting 1,000 token sentences containing prepositional phrases headed by the preposition *to*. Of these, approximately 500 were tokens for which the procedure guessed verb attachment and another 500 were tokens for which the procedure guessed noun attachment. Two judges were asked to disambiguate the token sentences. The judges were instructed to assign support verb constructions to noun attachment (see (3.29)), since they are not prototypical instances of the 'NP1 V NP2 *to* NP3' frame. In fact, the support verb and the noun form a semantic unit which in turn selects a prepositional complement.

A high precision was achieved in detecting both verb and noun attachment for *to*-PPs (see Table 3.7 where the mean of the classification of both judges is given). 74.4% of the tokens for which the procedure guessed verb attachment were genuine instances of verb attachment. Agreement on the classification task was also good with $K = .78$ ($N = 494$, $k = 2$). The procedure received a precision of 80.0% in detecting noun attachment. The two judges reached $K = .80$ ($N = 500$, $k = 2$). A fairly small number of support verb constructions (16.5% in total) was contained in the 1,000 sample sentences. Types of support verbs attested in the data are shown in (3.29). The method classified 13.9% of the overall support verb constructions as instances of verb attachment and the remaining 86.1% as instances of noun attachment.

(3.29) a. bear {relation, resemblance, witness} to
b. give {consent, consideration, effect, rise, way} to
c. make {appeal, contribution, difference, improvement, reference, rise, visit} to
d. {bring, draw, pay} attention to
e. pay {compliments, visit} to
f. pose threat to

Recall that in order to consider tokens with *to*-PPs as instances of the 'NP1 V NP2 *to* NP3' frame, the prepositional phrase must be the argument of the verb to which it is attached (see requirement 2 in Section 3.2.2.4). Although the procedure is only guessing attachment, it turns out that it is quite good in detecting verbs which subcategorize for a noun and a prepositional phrase headed by *to*. The judges were further asked to decide which of the sample tokens assigned to verb attachment by the LA-score were instances of the 'NP1 V NP2 *to* NP3' frame. As shown in Table 3.7, for 73.4% of the sample tokens, the PP headed by *to* was an argument of the verb. Only 26.6% of PPs were adjuncts attached either to verb or the noun. The judges

Table 3.8: Precision of the lexical association procedure and attachment site of *for*-PPs

	Precision	K
Verb Attach-for	73.6%	.85
Noun Attach-to	36.0%	.88
Attach-for	73.9%	.90

reached an agreement of $K = .90$ ($N = 494$, $k = 2$)

Tokens containing prepositional phrases headed by *for* were evaluated similarly. A random sample of approximately 1,130 tokens was selected, containing 630 tokens for which the procedure guessed verb attachment and 500 for which the procedure guessed noun attachment. Again two judges evaluated the procedure's performance by disambiguating the sample tokens. A major difficulty in disambiguating tokens containing the preposition *for*, also pointed out by Hindle and Rooth (1993), is that in several cases the preposition semantically licenses both attachment sites. The examples in (3.30) illustrate this point. We decided to make an attachment choice in all cases. In cases of indeterminacy with respect to the attachment site, the judges were instructed to choose the more likely site according to their intuitions.

(3.30) a. Although swap space maybe increased it is a drastic solution simply to obtain a parse for a single sentence.

b. Honey's work also presents difficulties for the viewer.

Although the procedure performs fairly well on detecting verb and noun attachment for *to*-PPs, this is not the case for *for*-PPs. As shown in Table 3.8 the procedure cannot distinguish instances of noun attachment from instances of verb attachment. A low precision of 36.0% was achieved in detecting instances of noun attachment. 64.0% of instances of verb attachment were misclassified as noun attachment. The judges reached a $K = .85$ ($N = 630$, $k = 2$) on judging verb attachment and $K = .88$ ($N = 500$, $k = 2$) on judging noun attachment.

There are several explanations as to why the lexical association score fails to distinguish noun attachment from verb attachment. One reason is the polysemy of the preposition *for*. As shown in (3.31) prepositional phrases headed by *for* can be temporal adjuncts (see sentence (3.31a)), purpose adjuncts (see example (3.31b)), benefactive, or causal adjuncts (see examples (3.31c) and (3.31d), respectively) and consequently can attach at various sites. Another difficulty is the semantic indeterminacy with respect to the site (verb or noun) licensing the attachment.

(3.31) a. He had felt sure when he left his hotel room that every eye in Saigon would be on him that evening because he was wearing a white tuxedo for the very first time in his young life.

b. You must not use official equipment for private purposes.

c. He replied that he wouldn't cut his hair for the Prime Minister.

d. They won't want to fire him for bad writing.

To further analyze the poor performance of the LA-score on this task, 500 tokens containing *for*-PPs were randomly selected from the parser's output and presented to judges who were asked to disambiguate the attachment site of the *for*-PP (see Table 3.8). Of these 73.9% were instances of verb attachment ($K = .90$, $N = 500$, $k = 2$), which indicates that verb attachments outnumber noun attachments for *for*-PPs, and therefore a higher precision can be achieved without applying the LA-score, but instead classifying all instances containing *for*-PPs as verb attachment.[7]

3.2.2.6. Benefactive PPs

Recall that in order to consider verbs attested in the 'NP1 V NP2 *for* NP3' frame as candidates for the benefactive alternation the *for*-PP must receive a benefactive interpretation (see requirement 3 in Section 3.2.2.4). Although surface syntactic cues can be important for deciding the attachment site of prepositional phrases, they provide no indication of the semantic role of the preposition in question. This is particularly the case for the preposition *for* which can have several roles, besides the benefactive.

Two judges discriminated benefactive from non-benefactive PPs for 500 tokens randomly selected from the parser's output. Only 18.5% ($K = .73$, $N = 500$, $k = 2$) of the sample contained benefactive PPs. This means that a hypothetical classifier that would consider all instances of the 'NP1 V NP2 *for* NP3' frame as benefactive would receive an accuracy of 18.5%. An analysis of the nouns headed by the preposition *for* revealed that 59.6% were animate (e.g., *student*, *burglar*, *animal*), 17.0% were collective (e.g., *community*, *charity*, *ministry*), 4.9% denoted locations (e.g., *Germany*, *region*, *center*), and the remaining 18.5% denoted events (e.g., *wedding*), artefacts (e.g., *house*), body parts (e.g., *back*), or actions (e.g., *prosecution*). Animate, collective, and location nouns accounted for 81.5% of the benefactive data.

We used the WordNet taxonomy (Miller et al. 1990) to recognize benefactive PPs. Recall from Chapter 2 that nouns in WordNet are organized into an inheritance system defined by hypernymic (superordinate) relations. Nouns are not contained in a single hierarchy; instead they are partitioned according to a set of semantic primitives which are treated as the unique roots of separate hierarchies. The 25 unique roots in WordNet are listed in Table 3.9. We compiled a "concept dictionary" from WordNet, where each entry consisted of the noun and the semantic primitive distinguishing each noun sense. Senses in WordNet are ordered by frequency, where frequency reflects the lexicographer's intuitions. A sample of the dictionary entries is given in Table 3.10. For example, the entry for the noun *gift* specifies that the word

[7] Spivey-Knowlton and Sedivy (1995) report a similar result for the preposition *with*. They extracted all sentences from the Brown corpus that contained a verb followed by a determiner, followed by a noun, followed by *with* (231 in total) and observed that VP attachments (62.0%) outnumbered NP attachments (38.0%).

Table 3.9: Semantic primitives for nouns in WordNet

act	*cognition*	*group*	*phenomenon*	*relation*
animal	*communication*	*location*	*plant*	*shape*
artefact	*event*	*motive*	*possession*	*state*
attribute	*feeling*	*object*	*process*	*substance*
body	*food*	*person*	*quantity*	*time*

Table 3.10: Sample entries from the WordNet concept dictionary

	Sense 1	Sense 2	Sense 3	Sense 4	Sense 5
mother	*person*	*substance*	*person*	*person*	*cognition*
gift	*possession*	*cognition*	*act*	———	———
alliance	*state*	*relation*	*group*	*communication*	*act*
charity	*group*	*attribute*	*act*	*plant*	*group*

has three senses, the primitive concept for Sense 1 is *possession*, for Sense 2 *cognition*, and for Sense 3 *act*.

We considered a *for*-PP to be benefactive if the noun headed by *for* was listed in the concept dictionary and the semantic primitive of its prime sense (Sense 1) was *person*, *animal*, *group*, or *location*.[8] PPs with head nouns not listed in the dictionary were considered benefactive only if their head nouns were proper names. Tokens containing personal, indefinite and reflexive pronouns were also considered benefactive (see cases (h) and (j) in Table 3.6).

We chose to consider only head nouns with prime concepts *person*, *animal*, *group* and *location* as indicators of benefactive PPs primarily based on the fact that these concepts account for 81.5% of the head nouns attested in the benefactive data. Although there is no hard and fast rule explicitly requiring that the complement noun of benefactive PPs is animate, in most cases *for*-PPs have to be animate in order to alternate with their ditransitive counterparts.[9] Contrast examples (3.33) and (3.34) below, where the alternation is acceptable in (3.33b)) and marked in (3.34b)). Quirk, Greenbaum, Leech, and Svartvik (1985) point out that this is because *magnolia tree* is not animate and therefore it does not qualify for the recipient role.

(3.33) a. I've found a place for Mrs Jones.
b. I've found Mrs Jones a place. (Quirk et al. 1985: 741)

(3.34) a. I've found a place for the magnolia tree.

[8] Another possibility would be to consider the *for*-PP benefactive if the majority of the head noun's senses is *person*, *animal*, *group*, or *location*. This is a less strict criterion than the one we applied.

[9] Strictly speaking, *for*-PPs which alternate with the ditransitive form must have a recipient interpretation. As shown in (3.32) deputive uses of *for*, where an action that was supposed to have been performed by one person is performed by a different person instead, to the benefit of the first person, do not have a ditransitive counterpart.

(3.32) a. John taught the syntax class for Mary.
b. *John taught Mary the syntax class. (Wechsler 1995: 83)

Table 3.11: Precision of detection of benefactive PPs using WordNet

	Precision	K
Benefactive	48.8%	.89
Non-Benefactive	90.9%	.94

b. ?I've found the magnolia tree a place. (Quirk et al. 1985: 741)

The process recognized 22,092 benefactive tokens. Two judges evaluated the outcome by judging 1,000 randomly selected tokens, of which 500 were classified as benefactive and 500 as non-benefactive. As shown in Table 3.11 the method achieved an average precision of 48.8% in detecting benefactive tokens. This is an improvement over considering all instances of the 'NP1 V NP2 *for* NP3' frame benefactive which yields an accuracy of 18.5%. Non-benefactive tokens were classified correctly with a combined precision of 90.9%. Agreement on classifying positive instances of benefactive PPs was $K = .89$ ($N = 500$, $k = 2$), whereas agreement on classifying negative instances was $K = .94$ ($N = 499$, $k = 2$).

3.2.2.7. Filtering

Filtering assesses how probable it is for a verb to be associated with a wrong frame. Erroneous frames can be the result of tagging errors, parsing mistakes, or further errors introduced by the heuristics and various procedures we used in order to guess syntactic structure in the first place (i.e., compound detection, PP attachment disambiguation, recognition of benefactive PPs).

A well known method for discarding erroneous frames is hypothesis testing on binomial frequency data. The method has been suggested by Brent (1993) and has been used for the acquisition of large-scale subcategorization dictionaries from corpora by Manning (1993) and Briscoe and Carroll (1997). The method works as follows: let p be an estimate of the probability that a verb token which does not occur with syntactic pattern s will nevertheless appear with s. Assume a verb v appears n times in a corpus, m times of which it occurs with a particular syntactic pattern s. Assume that v does not have the frame s, (null hypothesis H_0). Show that if H_0 was true, the observed pattern of co-occurrence of v with s would be extremely unlikely. Reject H_0 if the sum in (3.35) is smaller than a given confidence level (e.g., .05). The value of the estimate p can be set either empirically (Manning 1993) or determined automatically (Brent 1993; Briscoe and Carroll 1997).

$$(3.35) \quad P(+m,n,p) = \sum_{i=m}^{n} \frac{n!}{m!(n-m)!} p^m (1-p)^{n-m}$$

The hypothesis testing approach usually assumes that a full-scale subcategorization dictionary is extracted from the corpus through shallow syntactic analysis. Several frames are obtained for a given verb, the correctness of which is assessed via (3.35). The syntactic analysis

usually concentrates on surface structure without taking semantic information into account. For example, both sentences (3.36a) and (3.36b) would be considered as instances of the frame 'V NP1 *for* NP2' without distinguishing the benefactive *for*-PP (see (3.36b)) from the non-benefactive one (see (3.36a)).

(3.36) a. More than a quarter of voters said they would vote Perot for President if he ran.
b. I voted Bill Clinton for you.

Only the frames characteristic of the alternation at hand were obtained from our analysis. There was no global corpus analysis and no complete set of frames was extracted for a given verb. Furthermore, the corpus tokens obtained from the parser were post-processed in order to eliminate parsing errors (e.g., compound noun detection, prepositional phrase attachment). Although shallow, our approach tried to discriminate tokens like (3.36b) (which have benefactive *for*-PPs and are therefore relevant for the benefactive alternation) from tokens like (3.36a). Finally, it is not entirely clear that hypothesis testing can accurately eliminate false positives from true positives. Briscoe and Carroll (1997) showed that the performance of the hypothesis testing filter is around chance for subcategorization frames with frequency less than 10 and that a simple heuristic which considers true positives all frames with frequency more than 10 would have produced similar results to the hypothesis testing filter. Given that we did not conduct a full-scale experiment for the acquisition of subcategorization frames, we adopted a more naive approach to filtering erroneous subcategorization frames.

Initially, we discarded verbs for which we had very little evidence (frame frequency = 1); secondly, we applied a cutoff with respect to relative frame frequency: the verb's acquired frame frequency was compared against its overall frequency in the BNC. Verbs whose frame frequency was lower than a relative frequency threshold were discarded. The threshold values varied from frame to frame but not from verb to verb and were tuned experimentally by taking into account how probable a given frame is. The probability of each frame was estimated from the COMLEX subcategorization dictionary (Grishman et al. 1994, see chapter 2 for details). This meant the threshold was higher for less frequent frames (e.g., the double object frame for which only 79 verbs are listed in COMLEX) and lower for more frequent ones. The threshold for the double object frame was .5, for the *to*-PP frame was .3, and for the *for*-PP structures .2.

3.2.3. Results

We acquired 145 verb types for the double object frame, 426 verb types for the 'V NP1 *to* NP2' frame and 962 for the 'V NP1 *for* NP2' frame. Membership in alternations was judged as follows: (a) a verb participates in the dative alternation if both the double object and the 'V NP1 *to* NP2' frame have been acquired and similarly, (b) a verb participates in the benefactive alternation if both the double object and the 'V NP1 *for* NP2' frame have been acquired.

Table 3.12: Verbs common in the corpus and Levin

Dative Alternation	
Alternating	allocate, allot, assign, award, bequeath, bring, cede, concede, drag, fax, feed, flick, give, grant, guarantee, hand, issue, lease, leave, lend, offer, owe, pass, pay, pose, preach, promise, push, quote, read, render, repay, sell, send, serve, ship, show, slip take, teach, tell, throw, toss, write, yield
V NP1 *to* NP2	carry, catapult, drive, extend, ferry, fly, haul, hoist, relay, tug
V NP1 NP2	ask, chuck, shoot
Benefactive Alternation	
Alternating	bake, build, buy, call, cast, choose, cook, dig, earn, fetch, find, fix, forge, gain, get, hire, keep knit, leave, make, play, pour, prepare, procure, reserve, run, save, secure, set, shoot, toss, win, write
V NP1 *for* NP2	arrange, assemble, carve, compile, design, develop, dig, gather, grind, sew
V NP1 NP2	boil

Table 3.12 shows a comparison of our results against Levin's (1993) list of verbs: rows 'V NP1 NP2', 'V NP1 *to* NP2', and 'V NP1 *for* NP2' contain verbs listed as alternating in Levin (1993) but for which only one frame was acquired. In Levin 115 verbs license the dative and 103 license the benefactive alternation. Of these we acquired 45 for the dative and 33 for the benefactive alternation (in both cases excluding verbs for which only one frame was acquired).

The dative and benefactive alternations were also acquired for 67 verbs which are not listed in Levin. Of these 12 correctly alternate (see the examples in (3.37)–(3.38) and Table 3.13), and 20 can appear in either frame but do not alternate (see Table 3.13). Consider for example the verb *appoint*. It can take both the double object and prepositional frames as shown in (3.39a) and (3.40a) below. However, *appoint* does not participate in the dative alternation (see (3.39b) and (3.40b)). The same is true for *show*. In sentence (3.41a) *show* appears with the benefactive PP *for their friends*; the ditransitive alternant of (3.41a) does not preserve the benefactive meaning (see sentence (3.41b)). Such erroneous pairings cannot be avoided since our method only focuses on shallow syntactic/semantic regularities and does not take selectional restrictions into account. For 24 verbs two frames were acquired but only one frame was correct (see row 'One-frame' in Table 3.13) and finally 10 verbs neither alternated nor had the acquired frames. (see row 'No-frame' in Table 3.13).

(3.37) a. He delivers a lecture to his wife on his curious business.
b. He delivers his wife a lecture on his curious business.

(3.38) a. A consumer group recently forced the withdrawal of a medical advertisement which invited doctors to prescribe tranquilizers for air traffic controllers.
b. A consumer group recently forced the withdrawal of a medical advertisement which invited doctors to prescribe air traffic controllers tranquilizers.

Table 3.13: Dative and benefactive verbs found in the corpus only

Dative Alternation	
Alternating	deliver, refuse, accord, allow, cause
Non-alternating	set, appoint, declare, fix, gain, make, deny, get, permit, proclaim
One-frame	report, introduce, join, unveil, gain, afford, secure, win
No-frame	celebrate, clinch, incorporate, precede, prevent, regulate, renounce
Benefactive Alternation	
Alternating	afford, bring, chuck, deliver, prescribe, spare, spoil
Non-alternating	allow, appoint, cause, consider, forgive, give, guarantee, offer, pay, proclaim, show
One-frame	clinch, congratulate, disguise, implement, incorporate, inspect, install, integrate, introduce, launch, regain, retrieve, unveil, declare promise, swap
No-frame	attend, denounce, resent

(3.39) a. The President can either appoint him Chancellor or dissolve the Bundestag.
b. *The President can either appoint Chancellor to him or dissolve the Bundestag.

(3.40) a. Yeltsin was given the authority to appoint people to local government.
b. *Yeltsin was given the authority to appoint local government people.

(3.41) a. Those who had video-recorders also showed films for their friends.
b. Those who had video-recorders also showed their friends films.

Levin (1993) assumes that verbs participating in the same alternations share certain meaning components. Accordingly, she defines semantic classes of verbs characteristic for a given alternation. Levin defines 10 semantic classes that undergo the dative alternation (GIVE verbs, verbs of FUTURE HAVING, BRING AND TAKE, SEND, SLIDE, CARRY, DRIVE, THROWING verbs, verbs of TRANSFER OF A MESSAGE, and verbs of INSTRUMENT OF COMMUNICATION) and five classes that license the benefactive alternation (BUILD, CREATE, PREPARE verbs, verbs of PERFORMANCE, and GET verbs).

We partitioned our data according to Levin's (1993) predefined classes for the dative and benefactive alternation. Figure 3.2 shows for each semantic class the number of verbs acquired from the corpus against the number of verbs listed in Levin. We have excluded from the comparison verbs listed in Levin (1993) with overall corpus frequency less than one per million.[10] Levin and the corpus approximate each other for verbs of FUTURE HAVING (e.g., *guarantee*), GIVE verbs (e.g., *sell*), verbs of TRANSFER OF A MESSAGE (e.g., *tell*), and BRING AND TAKE verbs (e.g., *bring*). The semantic classes of GET verbs (e.g., *buy*) and PREPARE verbs (e.g., *pour*) are also represented in the corpus, in contrast to DRIVE and SLIDE verbs

[10] For the dative alternation these verbs are *bunt, bus, email, extend, heft, modem, netmail, punt, radio, satellite, schlep, semaphore, shuttle, telecast, telegraph, telex, tote, truck*, and *wireless*. For the benefactive alternation these verbs are *chisel, crochet*, and *hum*.

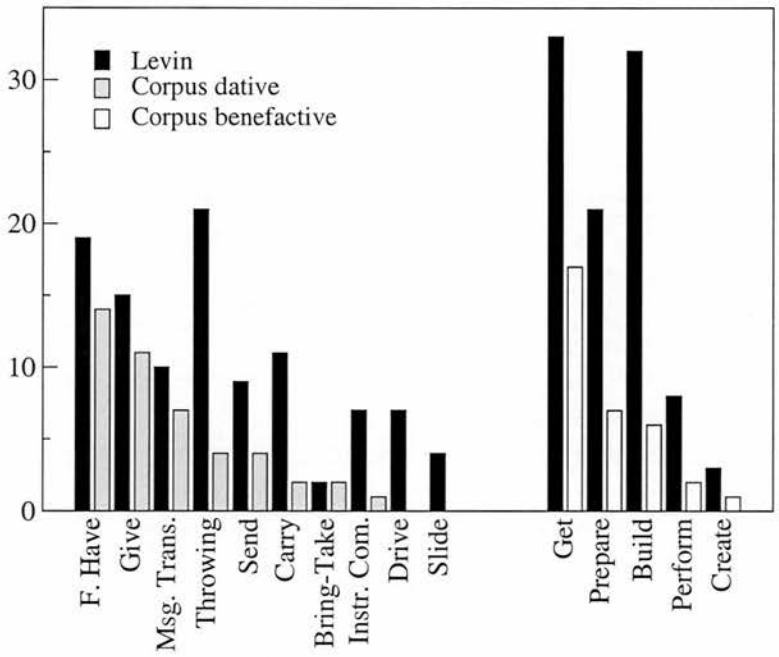

Figure 3.2: Semantic classes for the dative and benefactive alternations

(e.g., *fly* and *bounce*, respectively) for which no instances were found.

Note that the corpus and Levin (1993) did not agree with respect to the most popular classes licensing the dative and benefactive alternations: THROWING (e.g., *toss*) and BUILD verbs (e.g., *carve*) are the biggest classes in Levin allowing the dative and benefactive alternations respectively, in contrast to FUTURE HAVING and GET verbs in the corpus. This can be explained by looking at the average corpus frequency of the verbs belonging to the semantic classes in question: FUTURE HAVING and GET verbs outnumber THROWING and BUILD verbs by a factor of two to one.

In the following sections we show how the derived frame type and token frequencies can be used to estimate how productive an alternation is for a given verb semantic class and how typical its members are.

3.2.3.1. Productivity

The relative productivity of an alternation for a semantic class can be estimated by calculating the ratio of acquired to possible verbs undergoing the alternation (Aronoff 1976; Briscoe and Copestake 1999):

$$P(acquired|class) = \frac{f(acquired, class)}{f(class)} \tag{3.42}$$

Table 3.14: Productivity and typicality estimates for the dative and benefactive alternation

Dative alternation							
Class	Total	Alt	Prod	AvTyp	StdDev	Min	Max
BRING-TAKE	2	2	1.00	.78	.05	.75	.82
F. HAVE	19	14	.73	.64	.24	.18	.91
GIVE	15	11	.73	.51	.21	.23	.91
MSG. TRANS.	10	7	.70	.48	.34	.03	.90
CARRY	11	2	.18	.90	.03	.88	.92
DRIVE	7	0					
THROWING	21	4	.31	.50	.38	.38	.59
SEND	9	4	.44	.62	.10	.55	.76
INSTR. COM.	7	1	.14	.52			
SLIDE	4	0					
Benefactive alternation							
Class	Total	Alt	Prod	AvTyp	StdDev	Min	Max
GET	33	17	.51	.47	.29	.06	.99
PREPARE	21	7	.33	.43	.27	.11	.88
BUILD	32	6	.19	.42	.11	.27	.55
PERFORM	8	2	.25	.50	.08	.44	.56
CREATE	3	1	.33	.57			

We express the productivity of an alternation for a given class as the number of verbs which were found in the corpus and are members of the class, $f(acquired, class)$, over $f(class)$, the total number of verbs which are listed in Levin as members of the class (Total). The productivity values (Prod) for both the dative and the benefactive alternation (Alt) are summarized in Table 3.14.

Note that the productivity score is sensitive to class size. Since the class of BRING-TAKE verbs contains only two members which were also found in the corpus, the productivity of the dative alternation for this class is estimated to be one. On the one hand, this is intuitively correct, as we would expect specialized classes to be more productive. On the other hand, the estimate might be misleading, since it is only based on types and does not take token frequency into account. Consider for instance the verb *take*. We have acquired 1,810 instances for the prepositional frame and 880 instances for the double object frame. These frequencies clearly indicate an imbalance between the two frames which in turn shows that the alternation is not highly typical for *take*. In the following section we suggest a way of taking token frequency into account.

The productivity estimates discussed here can be potentially useful for treating lexical rules probabilistically, and for quantifying the degree to which language users are willing to apply a rule in order to produce a novel form (Briscoe and Copestake 1999). Linguistic theories such as Head-Driven Phrase Structure Grammar (HPSG, Pollard and Sag 1994), Lexical-Functional Grammar (LFG, Bresnan 2000) and Categorial Grammar (Steedman 2000) rely heavily on the application of lexical rules. Standard non-probabilistic formulations of lexical

rules assume that lexical rules are fully productive (i.e., the apply for any item matching their input specification). However, the semi-productivity of lexical rules has been known at least since Jackendoff (1975). The need for semi-productive lexical rules has been mainly advocated for morphology (Aronoff 1976), but also for syntax (Briscoe and Copestake 1995) and semantics (Briscoe and Copestake 1999; Copestake and Lascarides 1997).

Briscoe and Copestake (1995, 1999) have proposed to treat semi-productive lexical rules probabilistically. They argue that "lexical rules are sensitive to type and token frequency effects which not only determine language users' assessments of the degree of acceptability of a derived form but also their willingness to apply a rule in producing a novel form" (Briscoe and Copestake 1999: 511). The work reported here is in the spirit of their proposal since it allows us to estimate probabilistically the productivity of lexical rules applying to diathesis alternations.

It is well known (Boguraev and Briscoe 1989; Levin 1993) that diathesis alternations are not uniformly attested for all verbs of a given class. Although *invent* and *dig* are semantically assigned the same class (i.e., they are both CREATE verbs, Levin 1993), the benefactive alternation applies only to the latter (see the examples in (3.43) and (3.44)). In the next section we propose a measure that quantifies how likely it is for a given verb to be attested with a certain frame by taking frame frequency into account. Finally, note that a single verb may undergo more than one alternation. The verb *write*, for instance, may undergo both the dative and benefactive alternation (as a MESSAGE TRANSFER and PERFORMANCE verb, respectively), as illustrated by the examples in (3.45)). In the absence of contextual information, the choice between these two options can be cast in terms of choosing the alternation with the highest productivity (note that the productivity value for MESSAGE TRANSFER verbs is .70, whereas the productivity value for PERFORMANCE verbs is .25, see Table 3.14).

(3.43) a. It was difficult to invent a song for a princess.
b. *It was difficult to invent a princess a song.

(3.44) a. Royle refused, so City and Swales, having dug a large hole for themselves, returned to Kendall and agreed to his terms.
b. Royle refused, so City and Swales, having dug themselves a large hole, returned to Kendall and agreed to his terms.

(3.45) a. John wrote him a friendly letter.
b. John wrote a friendly letter to him.
c. John wrote a friendly letter for him.

3.2.3.2. Typicality

Estimating the productivity of an alternation for a given class does not incorporate information about the frequency of the verbs undergoing the alternation. We propose to use frequency data

Table 3.15: Typicality estimates for the dative and benefactive prepositional frames

GIVE	Typ	PREPARE	Typ
feed	.29	boil	.00
give	.23	cook	.34
lease	.90	fix	.48
leave	.65	pour	.11
lend	.65	prepare	.88
pay	.47	run	.62
render	.30	set	.42
repay	.58	toss	.13
sell	.76		
serve	.40		
write	.40		

to quantify the typicality of a verb or a verb class for a given alternation. The underlying assumption is that a verb is typical for an alternation if it is equally frequent in both frames which are characteristic for the alternation. Thus, the typicality of a verb can be defined as the conditional probability of the frame f given the verb v.

$$P(f|v) = \frac{f(f,v)}{\sum_i f(f_i,v)} \tag{3.46}$$

We calculate $P(f|v)$ by dividing $f(f,v)$, the number of times a verb v was attested in the corpus with the alternating frame f, by $\sum_i f(f_i,v)$, the overall number of times the verb was attested. In our case a verb has two alternating frames, hence $P(f|v)$ is close to .5 for typical verbs (i.e., verbs with balanced frequencies) and close to either 0 or 1 for peripheral verbs (i.e., verbs with imbalanced frequencies), depending on their preferred frame. Consider the verb *pay* as an example: 1,199 instances of *pay* were found, of which 567 were instances of the dative prepositional frame. By dividing the latter by the former we can see that *pay* is highly typical of the dative alternation: its typicality score for the prepositional frame is .47. Table 3.15 shows the typicality values for and GIVE and PREPARE verbs. These values have been calculated for the prepositional frames 'V NP1 *to* NP2' and 'V NP1 *for* NP2', respectively.

By taking the average of $P(f|v)$ for all verbs which undergo the alternation and belong to the same semantic class, we can estimate how typical this class is for the alternation. Table 3.14 illustrates the average typicality (AvTyp) for each semantic class for the dative and benefactive alternations. The standard deviation of the mean (StdDev) and its minimum (Min) and maximum (Max) values are also reported. For the dative alternation, the most typical classes are GIVE, MESSAGE TRANSFER, and THROWING verbs, whereas the most peripheral are CARRY verbs (e.g., *shove*). For the benefactive alternation, GET and PERFORMANCE verbs (e.g., *buy* and *sing*, respectively) are the most typical, BUILD and PREPARE verbs are

relatively typical (e.g., *carve* and *bake*, respectively), whereas CREATE verbs (e.g., *compose*) are relatively peripheral, which seems intuitively correct.

3.2.4. Discussion

Experiment 1 explored the degree to which diathesis alternations can be recognized in corpus data via shallow syntactic processing. Although Levin (1993) lists a greater number of dative and benefactive verbs than was actually acquired from the corpus, the results indicate that both the dative and benefactive alternation are well attested in the BNC. A heuristic which classifies verbs as alternating simply if they are attested in the corpus with both related frames performs adequately in terms of discovering not only verbs which are known to alternate (i.e., are listed in Levin) but also novel verbs (i.e., verbs which do alternate but are not listed in Levin).

The acquired verbs and their frame frequencies can be used to estimate the productivity of an alternation for a given semantic class and the typicality of its members. The productivity and typicality estimates make explicit predictions about word use. For instance, classes with low productivity and typicality values are less likely to exhibit the alternation in comparison to classes with high values. We test this prediction in Experiment 4.

Despite the fact that a shallow approach seems to work relatively well, Experiment 1 demonstrated some of the key difficulties in acquiring diathesis alternations from corpora. The acquisition of the benefactive 'V NP1 *for* NP2' frame clearly illustrates that surface structure cannot distinguish benefactive from non-benefactive PPs. Levin's (1993) description of the argument structure of various verbs goes beyond the simple listing of their subcategorization. The classification of verb argument structure into different types of alternations exploits information about the thematic roles of verbal arguments and their meaning which is not available to our heuristic approach. Nevertheless, a shallow semantic approach which exploits the WordNet taxonomy performs adequately at recognizing instances of benefactive PPs. We exploit WordNet's semantic information again in Experiment 2, where we examine the conative alternation, whose prepositional frame variant poses similar difficulties.

Experiment 1 focused on alternations which do not involve changes in the number of arguments of the alternating verbs. Both benefactive and dative verbs, irrespectively of their surface syntactic realization (i.e., prepositional or not) require two objects. Furthermore, the changes in the argument structure are not accompanied by changes in meaning. In Experiment 2 we use the same methodology to look at the conative alternation which involves a change in the number of arguments of the alternating verbs and also changes in meaning.

3.3. Experiment 2: The Conative Alternation

3.3.1. Introduction

The conative alternation involves a change in the verb's transitivity (Levin 1993). The phenomenon in English is illustrated in (3.10) repeated here as (3.47).

(3.47) a. Paula hit the fence.
b. Paula hit at the fence.

(3.48) a. The mouse nibbled the cheese.
b. The mouse nibbled at/on the cheese. (Levin 1993: 42)

(3.49) a. Janet broke the bread.
b. *Janet broke at the bread. (Levin 1993: 41)

Verbs undergoing this alternation can be attested either as transitive (see examples (3.47a) and (3.48a)) or as intransitive where the object of the transitive variant is the object of a prepositional phrase headed by the preposition *at* and sometimes by the preposition *on* or *onto* (see sentences (3.47b) and (3.48b)). For the verbs that undergo the conative alternation the change in the realization of their argument structure (transitive versus intransitive) is accompanied by a change in meaning. The use of the verb in the intransitive variant describes an *attempted action* without necessarily entailing that the action was carried out (Dixon 1991; Levin 1993). Consider example (3.47): sentence (3.47a) implies that Paula delivered a hit to the fence, whereas sentence (3.47b) implies that Paula tried to hit the fence, but the aim was not achieved. Similarly, example (3.48a) implies that the mouse bit the cheese, i.e., a piece was bitten out of the cheese, and then chewed and swallowed, whereas example (3.48b) does not imply that a portion of cheese was eaten.

Levin (1993) lists 79 verbs which undergo the alternation and 147 verbs for which the prepositional variant is not possible (see example (3.49)). Table 3.16 gives a sample of alternating and non-alternating verbs taken from Levin.

In the following section we describe and evaluate the set of automatic methods we used to acquire verbs undergoing the conative alternation. We present our results in Section 3.3.3. Section 3.3.3.1 discusses how the derived type and token verb frame frequencies can be used to estimate the productivity of the alternation (i.e., whether verbs of a certain semantic class are more likely to alternate than others) and the typicality of its members (i.e., how likely is it for a verb to alternate given that it is attested in both frame alternants).

3.3.2. Method

3.3.2.1. Acquisition

The experimental framework was the same as in Experiment 1. Tokens characteristic for the conative alternation were extracted from a part-of-speech tagged and lemmatized version of

Table 3.16: Alternating and non-alternating verbs for the conative alternation

Alternating	Non-alternating
bite	belt
cut	bend
dig	cane
eat	chip
hit	cube
kick	dice
nib	slice
peck	spank
press	touch
stab	whisk

the BNC. Surface syntactic structure was again identified using Gsearch (Corley et al. 2001).

We used Gsearch to extract tokens matching the patterns 'V NP' and 'V *at* NP' corresponding to the transitive and intransitive variant of the alternation, respectively. In this experiment we ignored potential conative PPs headed by *on* or *onto*. Examples of the parser's output are given in (3.50)–(3.55). Although there are cases where Gsearch produces the right parse (see (3.50)), the parser wrongly identifies as instances of transitive verbs verbs taking sentential complements (see (3.51)), ditransitive verbs (see (3.52)), copula verbs (see (3.53)), and verbs followed by NPs which are not arguments but modifiers (see (3.54)). The examples in (3.55) are misparsed due to tagging errors: in sentence (3.55a) the gerund *praying* is tagged as a noun, resulting in an analysis where *praying* is identified as the object of *keep*, and in (3.55b) the verb *fly* is mistagged as a noun, leading to an analysis where the sequence *time fly* is identified as the object of *help*.

(3.50) a. Donald [$_V$ kicked] [$_{NP}$ the kettle] in a spasm of impatience.
b. The vet left them in the staff room while they tried to [$_V$ discuss] [$_{NP}$ Lizzie's future].

(3.51) a. When I was at school I hated it, but when I left I [$_V$ wished] [$_{NP}$ I] were back there.
b. Japanese suppliers [$_V$ assume] [$_{NP}$ they] will lose new orders to American rivals for political and protective reasons.
c. He [$_V$ thinks] [$_{NP}$ drugs] are evil.
d. She [$_V$ told] [$_{NP}$ me] that she was eight years old.

(3.52) a. Baudelaire was seeking to [$_V$ give] [$_{NP}$ new life] to a decayed literary genre.
b. I [$_V$ told] [$_{NP}$ you] a lie about that.

(3.53) a. To others it will [$_V$ appear] [$_{NP}$ an example] of the chapel 's commitment to their beliefs.
b. What [$_V$ seemed] [$_{NP}$ a difference of degree] turns out to be one of a kind.

(3.54) a. I cannot [$_V$ spend] [$_{NP}$ all day] with her.
b. A road for motorists [$_V$ continues] [$_{NP}$ east] from there.

(3.55) a. Please [$_V$ keep] [$_{NP}$ praying] that I will master the German ways of doing things.
b. And to [$_V$ help] [$_{NP}$ time fly] during their brief stay, there was a radio to listen to even magazines to read.

Although in most cases Gsearch produces the right parse for the 'V *at* NP' structure (see the sentences in (3.56)), it cannot distinguish between arguments and adjuncts. Consider the examples in (3.57): both verbs (*look* and *munch*) are followed by a PP headed by the preposition *at*. However, in example (3.57a), *look* is transitive followed by its object, the PP *at the display of cakes*, whereas in (3.57b) *munch* is intransitive modified by the PP *at the thin grass*. Furthermore, the parser cannot distinguish between different types of adjuncts. In the examples in (3.58) the preposition *at* is part of the fixed expressions *at the expense of* and *at the heart of*. In sentences (3.59a) and (3.59b) the PPs *at the scene* and *at Barcelona* are locative. In sentence (3.59c) the PP *at regular time-intervals* is temporal, whereas in (3.59d) the PP *at £3.1billion* receives a scalar interpretation. It is only in (3.59e) where the PP *at his chin* indicates the direction or goal of the action expressed by the verb it modifies (i.e., *scratch*) and is therefore relevant for the conative alternation.

(3.56) a. When he finished, he [$_V$ kicked] [$_{PP}$ at the door].
b. He [$_V$ bit] [$_{PP}$ at his lip] and swallowed again.

(3.57) a. He turned and [$_V$ looked] [$_{PP}$ at the display of cakes] on the long table.
b. Their horses [$_V$ munch] [$_{PP}$ at the thin grass] in a desultory fashion.

(3.58) a. A privileged caste which [$_V$ lives] [$_{PP}$ at the expense of society] as a whole.
b. The job description [$_V$ lies] [$_{PP}$ at the heart of good recruitment] and selection practice.

(3.59) a. The first person to [$_V$ arrive] [$_{PP}$ at the scene] was a flying instructor.
b. Simon stepped in after they [$_V$ met] [$_{PP}$ at Barcelona].
c. All syllables, whether stressed or unstressed, tend to [$_V$ occur] [$_{PP}$ at regular time-intervals].
d. The parent company, Racal Electronics, is currently [$_V$ valued] [$_{PP}$ at £3.1billion]
e. Rex [$_V$ scratched] [$_{PP}$ at his chin].

We identified erroneous subcategorization frames for tokens matching the 'V NP' pattern (see (3.51)–(3.55)) using the linguistic heuristics detailed in the following section. We attempted to isolate verbs followed by PPs which receive a conative interpretation (see (3.56) and (3.59d)) by combining linguistic heuristics and information provided by the WordNet taxonomy (Miller and Charles 1991).

3.3.2.2. Guessing the transitive frame

We applied several heuristics to the parser's output in order to distinguish tokens which are erroneous matches (false positives) from tokens which are genuine instances of the transitive frame (true positives). We partitioned again the heuristics in two classes: the 'Reject' heuristics identified false positives, whereas the 'Accept' heuristic identified true positives.

1. **Reject** if the verb is followed by a pronoun and the pronoun's case is nominative (see examples (3.51a,b)).
2. **Reject** if cue's surface structure is either 'V NP V' or 'V NP COMP'[11] (see examples (3.51c) and (3.51d), respectively).
3. **Reject** if the verb has been previously attested in the double object frame (see the sentences in (3.52)).
4. **Reject** if the verb is a copula verb (e.g., *appear, feel, remain, rest, seem, smell*) (see the sentences in (3.53)).
5. **Reject** if the verb is followed by noun phrase which is a temporal or locative modifier (see examples (3.54)).
6. **Accept** in all other cases.

Heuristics 1 and 2 apply on the surface structure of the parser's output. They both eliminate instances of verbs subcategorizing for sentential complements. Heuristic 1 handles transitive verbs (e.g., *wish, hope*), whereas heuristic 2 handles ditransitive verbs (e.g., *tell, think*). Heuristic 3 rejects tokens which have been previously attested in the double object frame (see Section 3.2.2.2). We are making here the unrealistic assumption that a verb cannot be both ditransitive (i.e., attested in the double object frame 'V NP NP') and transitive (i.e., attested in the frame 'V NP'). This is clearly not the case for *consider* which takes both a ditransitive and a double object frame (see examples (3.60a) and (3.60b), respectively).

(3.60) a. General Salan's punishment to life imprisonment in 1968 so angered President de Gaulle that he [$_V$ considered] [$_{NP}$ resignation].
b. John [$_V$ considers] [$_{NP}$ himself] [$_{NP}$ a jealous man].
c. Or you can [$_V$ forget] [$_{NP}$ the week] as a unit of work altogether.

Heuristic 4 eliminates copula verbs. A list of copula verbs was obtained from Quirk et al. (1985). Tokens with verbs contained in the list were discarded. Finally, heuristic 5 eliminates tokens with verbs followed by nominal modifiers. A list of nominal modifiers was manually

[11] Here COMP represents complementizers (e.g., *that, whether, if*).

created. The list contained the days of the week, the months and seasons of the year, and temporal expressions containing the words *year*, *month*, *week*, *day*, *morning*, *afternoon*, *evening*, *night*, *weekend*, *moment*, *hour*, and *minute* in a variety of syntactic realizations examples of which are shown in (3.61). Tokens with complements matching the expressions in the list were discarded. Heuristic 5 also relies on the simplifying assumption that expressions like the ones given in (3.61) are more likely to be modifiers than complements (however, see sentence (3.60c) for a counterexample).

(3.61) a. this|every|each|all (next|last) year
(the) next|last year
those|these|the (next|last) years
b. this|every|each moment

3.3.2.3. Guessing the intransitive frame

Although the parses obtained for the 'V *at* NP' pattern were in most cases structurally correct, Gsearch provides no cues as to whether the PP following the verb is its complement or adjunct. Even if we could make this distinction, the parser provides no information with respect to the semantic role of the preposition *at*. As shown in (3.59) the preposition *at* can have several meanings, besides the conative.

Two judges discriminated conative from non-conative PPs for 500 tokens randomly selected from the parser's output. Only 3.6% ($K = .82$, $N = 500$, $k = 2$) of the PPs contained in the sample were conative. This means that a hypothetical procedure which classifies all instances of the 'V *at* NP' frame as conative would receive an accuracy of 3.6%. Table 3.17 shows the distribution of the various types of *at*-PPs in the sample. Note that only 2.2% of the tokens included in the sample are misparsed. Furthermore, 40.8% of the tokens are instances of transitive verbs, whereas 45.2% of the tokens are instances of intransitive verbs modified by a locative, temporal, or modal adjunct. Of the verbs which subcategorize for an *at*-PP in the sample (16 in total), only a few take direct object NPs and therefore can be potentially mistaken as participating in the conative alternation. Table 3.18 shows the distribution of the transitive verbs found in the sample. Of these verbs only *shout*, *point*, *wave*, *scream* and *excel* also subcategorize for an NP object (see the examples in (3.62)). This is not the case for the intransitive verbs which are modified by locative, temporal and modal PPs: 80.0% of these verbs (122 in total) can be attested in the transitive frame and can be consequently mistaken as undergoing the conative alternation.

Table 3.17: Sample distribution of *at*-PPs

Interpretation	Tokens	Distribution
Complement	204	40.8%
Locative	151	30.2%
Temporal	52	10.4%
Target	35	7.0%
Modal	23	4.6%
Conative	18	3.6%
In reaction to	5	1.0%
Scalar	1	.2%
Misparses	11	2.2%

Table 3.18: Sample distribution of transitive verbs

Verb	Frequency	Verb	Frequency
look	137	glare	2
stare	26	gaze	2
glance	16	wonder	1
smile	8	wave	1
peer	5	scream	1
laugh	4	goggle	1
shout	3	glower	1
point	2	excel	1

(3.62) a. Hundreds of thousands shouted his name during the demonstrations at Tiananmen.
b. She pointed her right toe.
c. He was shouting and waving his arms.
d. Death screamed a curse in his cold crypt voice.
e. He excelled his inherited duties by outliving the three heiresses he married.

We eliminated tokens containing temporal, locative and modal PPs as follows: under the assumption that prepositional phrases which are frequently attested adjacent to the verb are fixed expressions and therefore cannot be conative, a list of the most frequent *at*-PPs (co-occurrence frequency ≥ 40) was compiled from the parser's output. Table 3.19 shows a random sample of these fixed expressions. Tokens with verbs followed by any of the expressions in the list were discarded (false positives). As a next step, tokens containing PPs whose head noun was either a number or a proper noun were also discarded under the assumption that such PPs are more likely to be temporal, modal, scalar, or locative rather than conative (see the examples in (3.63)). For the remaining tokens the WordNet concept dictionary described in Section 3.2.2.6 was used to identify primarily temporal and locative PPs. Tokens were considered locative or temporal only if the head noun of the *at*-PP was listed in WordNet and the prime meaning of its first sense was either temporal or locative.

(3.63) a. However, the overstretched executive still made £1.3m profit when he sold at 339p.
b. Her chauffeur was caught driving at 90mph in a 60mph zone.
c. In a previous journalistic incarnation, I worked at Westminster.

Note that we did not attempt to distinguish verbs which are transitive and their object is a PP headed by *at* (e.g., *look*) from genuinely intransitive verbs which are modified by conative *at*-PPs (e.g., *hit*). This is a considerably harder task for the heuristic approach advocated here which is knowledge-poor and relies on a chunk grammar to produce a partially parsed version of the corpus.

Table 3.19: Random sample of fixed expressions with their corpus frequencies

Frequency	Expression
1816	at times
360	at a rate
276	at hand
218	at this age
167	at a pace
136	at that price
98	at regular intervals
85	at the outset
66	at will
64	at the last minute

Table 3.20: Precision of the heuristics for the transitive frame and conative PPs

	Precision	*K*
Accept V NP	89.3%	.72
Reject V NP	93.0%	.75
Accept V *at* PP	73.8%	.80
Reject V *at* PP	74.9%	.78

3.3.2.4. Evaluation

The heuristics employed for guessing the transitive frame were evaluated by randomly selecting 500 tokens which were accepted as instances of the transitive frame and 500 tokens which were rejected as instances of the transitive frame. Similarly, the procedure developed for distinguishing conative from non-conative tokens was evaluated by randomly selecting 500 tokens which were accepted either as instances of transitive verbs or as instances of conative *at*-PPs and 500 tokens which were rejected as instances of intransitive verbs modified by non-conative *at*-PPs. Two judges decided whether the tokens were classified correctly. For the transitive frame the judges were asked to decide whether corpus tokens were instances of the transitive frame or not. Given that no process was developed for distinguishing conative PPs from PPs which are arguments, the judges were instructed to consider the latter as true positives. The heuristics' average precision as well as the inter-judge agreement are summarized in Table 3.20.

The heuristics achieved a high combined accuracy (89.3%) in identifying tokens which are instances of the transitive frame. A somewhat higher accuracy of 93.0% was obtained for the heuristics which discarded false positives (see Table 3.20). The procedure developed for the prepositional frame achieved a combined accuracy of 74.9% in recognizing non-conative PPs. Conative PPs and transitive verbs were identified correctly at 73.8% (see Table 3.20). These accuracy figures are good given the heuristic approach and the sparsity of conative PPs (recall that only 3.6% of *at*-PPs are conative).

3.3.2.5. Filtering

As in Experiment 1 we assessed how probable it is for a verb to be associated with a wrong frame. We first discarded verbs for which we had very little evidence (frame frequency = 1). Second, we applied a relative frequency threshold which varied from frame to frame but not from verb to verb: verbs with relative frame frequency below the threshold were discarded. Although the thresholds for the transitive and the prepositional frame were tuned experimentally, an initial estimate was obtained from the COMLEX subcategorization dictionary (Grishman et al. 1994). 14.0% of the verbs listed in COMLEX are transitive, whereas only .7% are either modified by or subcategorize for *at*-PPs. This means that it is approximately 20 times more likely for a verb to subcategorize for an NP than for an *at*-PP. Consequently a higher threshold was established for verbs acquired with the *at*-PP frame than for verbs acquired with the NP frame. The procedure yielded 265 verbs with the *at*-PP frame and 6,508 verbs with the NP frame.

3.3.3. Results

A verb was classified as undergoing the conative alternation if it had been found in the corpus with both the NP and *at*-PP frame. According to this simple criterion 101 verbs were found that license the conative alternation. Of these verbs, 36 undergo the alternation and were found both in Levin (1993) and the corpus, whereas 66 were found in the corpus only. Table 3.21 shows a comparison of the verbs found in the corpus against Levin's list of verbs; rows 'V NP' and 'V *at* NP' contain verbs listed as alternating in Levin's but for which we acquired only one frame. In Levin 79 verbs license the conative alternation. Note that 45.6% of these verbs were attested in the corpus in both frame alternants ('V NP' and 'V *at* NP') and furthermore, that for most verbs (74 out of 79) the transitive frame was acquired.

Out of the 65 verbs that were found in the corpus only with both the transitive and the prepositional frame, 23 undergo the conative alternation, 25 verbs can appear in either frame but do not alternate (in this case, the *at*-PP is either a verbal argument or receives a goal interpretation), and 17 verbs can appear only in one of the acquired frames (in this case, the *at*-PP receives a temporal, locative, or modal interpretation). These verbs are displayed in Table 3.22 and partitioned in three categories: alternating, non-alternating, and one-frame.

The procedure identified a large number of manner of speaking verbs (e.g., *bellow*, *bawl*, *croak*, *hurl*, *shout*, *shriek*, *preach*, *spit*). Although these verbs do not license the conative alternation (see non-alternating verbs in Table 3.22), they display a systematic alternation between the transitive 'V NP' and the intransitive 'V *at* NP' frame (see the examples in (3.64)–(3.66)). The 'V NP' variant is the so called reaction object construction where manner of speaking verbs (and verbs of gestures and signs) take objects that express a reaction (Levin 1993) (for instance, sentence (3.64a) can be paraphrased as "The artillery Colonel expressed his orders by

Table 3.21: Conative verbs common in corpus and Levin

Conative Alternation	
Alternating	bite, chew, claw, dab, gnaw, hack, hammer, heave, jab, kick, knock, lash, lick, munch, nibble, paw, peck, poke, pound, prick, rap, rub, scratch, shoot, sip, slash, snip, squirt, stab, strike, suck, swipe, tap, tug, whack, yank
V NP	bang, bash, batter, beat, butt, chomp, clip, crunch, cut, dig, draw, drink, drum, eat, hit, jerk, pick, pierce, press, pull, punch, push, saw, scrape, shove, slap, slug, slurp, smack, smash, splash, spray, stick, swab, swat, tamp, thrust, thump
V NP *at* PP	chip

Table 3.22: Conative verbs found in corpus only

Conative Alternation	
Alternating	chafe, clutch, flap, flick, grab, grasp, grope, guess, gulp, lap, mouth, nose, pluck, prod, pummel, scrub, snap, snatch, sniff, swig, target, tear, wrench
Non-alternating	anger, bawl, bellow, boggle, croak, fire, flash, fling, gag, hurl, nag, niggle, parade, preach, scoff, shout, shriek, slog, spit, spook, spurn, stop, tense, throw, wave
One-frame	angle, branch, crystallize, dock, improvise, intersect, key, lecture, parade, retail, slog, steam, stop, subtend, throw, waken, zip

bellowing"). In the prepositional variant the preposition *at* receives a goal interpretation; more specifically, the action expressed by the verb is directed toward the complement of the *at*-PP.

The procedure also identified verbs with both frame alternants (e.g., *boggle*, *gag*, *tense*, *spook*, *anger*) for which the prepositional variant means "in reaction to" (see the examples in (3.67)–(3.68)). We also identified verbs such as *flash*, *fire*, *fling* and *wave*: these verbs have a transitive and intransitive frame; as intransitives they are modified by directive *at*-PPs (see examples (3.69)–(3.70)). Finally, the verb *scoff* illustrates the limitations of the approach: the verb has two senses, when attested with the 'V *at* NP' frame it means "dismiss", whereas when attested with the 'V NP' frame it means "eat quickly" (see example (3.71)). This is clearly a case where our underlying assumption (i.e., that surface syntactic structure can provide cues about meaning) does not hold.

The procedure also acquired verbs for which the prepositional frame has either a locative, temporal, or modal interpretation. These verbs are listed in Table 3.22 under the heading 'One-frame'. These verbs differ from the non-alternating verbs in that they do not show a productive (even though not conative) alternation between the 'V NP' and 'V *at* PP' frame (see the examples in (3.72)).

(3.64) a. The artillery Colonel bellowed his own orders.
b. Cranston bellowed at the taverner.
(3.65) a. They croak a furtive greeting.
b. Irina croaked at her.
(3.66) a. Divorcee Dianne Wiest's fruity teenage offspring barely communicates with her except to hurl insults.
b. A mob of young Catholics was waiting at Cromac Square, armed with a good supply of bricks and metal objects to hurl at the marchers.
(3.67) a. He tensed his chest demonstratively.
b. Harry had tensed at the criticism.
(3.68) a. They have even spooked a couple of grizzly bears.
b. He spooked at some seemingly insignificant object.
(3.69) a. Howard waved his hand.
b. I waved at the receptionist.
(3.70) a. Joyce fired his pistol and the battle had begun.
b. I pulled my gun from my coat and fired at him.
(3.71) a. I never complained as I went off to watch West Ham that afternoon, having scoffed his portion of chips.
b. He scoffs at my mother sometimes.
(3.72) a. He lectured me like a schoolmaster.
b. He was exceptionally nice and he used to lecture at the colleges.

Levin (1993) defines eight semantic classes of verbs that license the conative alternation: HIT verbs (e.g., *beat, kick*), SWAT VERBS (e.g., *bite, paw*), POKE verbs (e.g., *pierce, stick*), CUT verbs (e.g., *chip, scrape*) SPRAY/LOAD verbs (e.g., *splash, rub*), PUSH/PULL verbs (e.g., *press, draw*), EAT verbs (e.g., *eat, drink*), and CHEW verbs (e.g., *nibble, munch*). Figure 3.3 compares the acquired verbs against Levin. The comparison excludes verbs listed in Levin with corpus frequency less than one per million.[12] Levin and the corpus approximate each other for CHEW and SWAT verbs. The class of HIT verbs, despite being the biggest class licensing the conative alternation both in Levin and the corpus, is relatively underrepresented in the corpus. The semantic classes of POKE verbs and SPRAY/LOAD verbs are also represented in the corpus, in contrast to EAT verbs for which no instances were found.

Let us now consider the verbs that were found to undergo the conative alternation in the corpus but not in Levin (1993). If Levin's hypothesis that the realization of the argument structure of a given verb is a direct reflection of its meaning is valid, then one would expect these verbs to fall under one of the eight semantic classes that license the conative alternation. The corpus data seems to support Levin's hypothesis. Of the 23 verbs that were found to alternate in the corpus only (see Table 3.22), four verbs (*flick, pluck, wrench, tear*) are members

[12]These verbs are *thwack, slug, hew, swab, chomp,* and *slurp.*

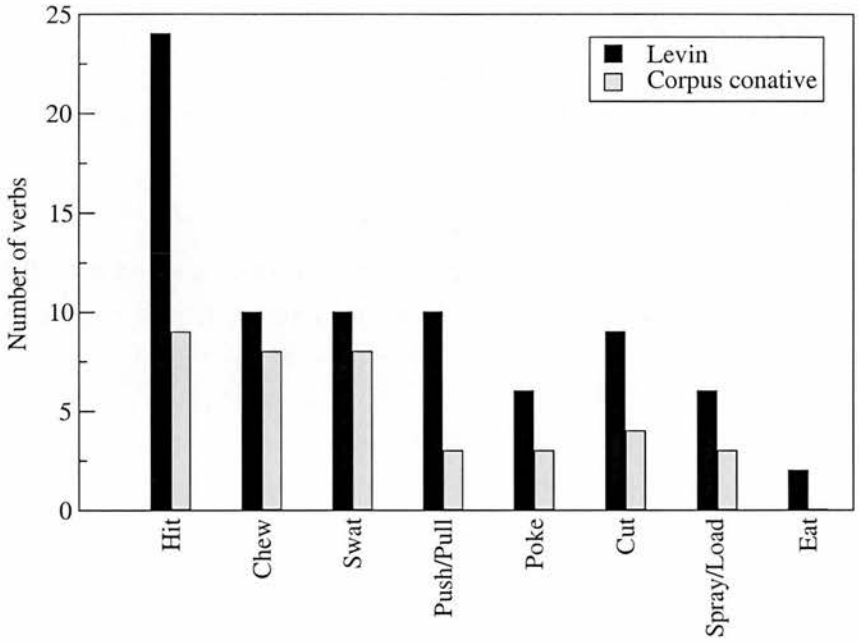

Figure 3.3: Semantic classes for the conative alternation

of the class of PUSH/PULL verbs (see the examples in (3.73) and (3.74)), three verbs (*prod, grope*, and *chafe*) are POKE verbs (see example (3.75)), two verbs (*gulp* and *swig*) are members of the DRINK class (see example (3.76)), two verbs (*grope* and *mouth*) are CHEW verbs (see example 3.77), two verbs (*pummel* and *flap*) are members of the HIT class (see example (3.78)) and one verb (*scrub*) is licensed by the SPRAY/LOAD class (see example (3.79)). Note that in most cases the alternants are attested with identical complements in the corpus (see examples (3.73)–(3.79)).

(3.73) a. A ruck was set up to the left of the Cambridge posts, when Smith flicked the ball out to the right.
b. He would flick at the ball with the outside of his left foot while leaning back looking at the sky.

(3.74) a. Alan Millet plucked his coat from the fender.
b. He plucked at his uniform coat as if meaning to tear it off.

(3.75) a. She put down the saucepan, prodded the pile of letters, tugged one out from farther down.
b. The man and I prodded at the pile of crap on the table.

(3.76) a. Bodo swigged his brandy.
b. Horrocks swigged at his brandy.

(3.77) a. She will occasionally mouth the clutch to remove any particles of debris.
b. Isobel smiled again as he mouthed at his fingers.

(3.78) a. Isabel pummeled the straw-filled pillow into a more comfortable shape.
b. Ashley pummeled at her pillow again.

(3.79) a. The minute she went in she was made to scrub the table – and it was white – scrub the floorboards – which were white.
b. Aunt Margaret scrubbed at her board for more space.

Finally, some verbs license the alternation without belonging to any of the eight semantic classes defined by Levin (1993). These are for instance OBTAIN verbs (*grab, snatch, snap*) and HOLD verbs (*grasp, clutch*) exemplified in (3.80) and (3.81), respectively.[13] This result is broadly compatible with Dixon (1991) who observes that the insertion of the preposition marks the fact that the actual identity of the object is of peripheral interest and that it is not affected by the action. Dixon (1991: 280) further notes that the prepositional variant places emphasis not on the effect of the activity on some specific object but rather on the subject's engaging activity.

(3.80) a. The boy snatched the ball from his hand and raced away.
b. He snatched at the ball inside the Bournemouth penalty area.

(3.81) a. I grasped the vase with both hands.
b. His hand grasped at the cold stone.

3.3.3.1. Productivity and Typicality

Recall from Section 3.2.3.1 that the relative productivity of an alternation for a semantic class can be estimated by calculating the ratio of acquired to possible verbs undergoing the alternation. Here we are only considering verbs common between Levin (1993) and the corpus. The productivity of values (Prod) for the conative alternation are summarized in Table 3.23. The conative alternation is highly productive for CHEW and SWAT verbs and fairly productive for POKE and SPRAY/LOAD verbs. Productivity estimates are complemented with typicality estimates which indicate how typical a verb is with respect to the alternation.

In Section 3.2.3.2 we defined typicality as the likelihood of a verb to be attested in one of the frame alternants. Verbs with typicality values close to 0 or 1 are atypical for the alternation in contrast to verbs with typicality values close to .5. These verbs have no strong preference for either frame and hence are more likely to alternate. The typicality values for HIT and SWAT verbs are shown in Table 3.24 (these values have been computed for the 'V *at* NP' frame). Note that the verbs in both classes strongly prefer the transitive frame (with the exception of *lash* and *paw*). This is in fact the case for all semantic classes licensing the conative alternation. The average typicality (AvTyp) for each semantic class is illustrated in Table 3.23 (the standard deviation of the mean (StdDev) and its minimum (Min) and maximum

[13] OBTAIN verbs undergo the benefactive alternation and HOLD verbs license the location subject alternation according to Levin (1993).

Table 3.23: Productivity and typicality estimates for the conative alternation

Class	Total	Alt	Prod	AvTyp	StdDev	Min	Max
HIT	24	9	.37	.08	.10	.03	.35
CHEW	10	8	.80	.18	.16	.01	.48
SWAT	10	8	.80	.12	.13	.01	.39
PUSH/PULL	10	3	.30	.14	.19	.03	.36
POKE	6	3	.50	.06	.03	.03	.09
CUT	9	4	.44	.07	.03	.04	.12
SPRAY/LOAD	6	3	.50	.11	.15	.02	.28
DRINK	2	0					

Table 3.24: Typicality estimates for the 'V *at* NP' frame

HIT verbs	Typ	SWAT verbs	Typ
hammer	.07	bite	.02
kick	.04	claw	.16
knock	.07	lick	.01
lash	.35	paw	.40
pound	.06	scratch	.04
rap	.06	shoot	.04
strike	.04	stab	.07
tap	.03	swipe	.22
whack	.03		

(Max) values are also reported). All conative classes contain atypical verbs: they systematically disprefer the prepositional frame. This is perhaps not surprising given that the transitive frame is far more frequent in the corpus than its prepositional variant (after discarding false positives, 2,331,071 tokens had the 'V NP' frame, whereas only 32,959 had the 'V *at* NP' frame). The typicality values indicate that conative verbs are less likely to alternate than for instance dative verbs for which higher typicality values were found (see Table 3.14 in Section 3.2.3.1).

3.3.4. Discussion

In Experiment 2 we further explored the surface cueing methodology introduced in Experiment 1 by acquiring frames characteristic of the conative alternation. The acquisition process faced the major difficulty of recognizing the conative frame. The difficulty arises from the fact that several verbs display a systematic alternation between the 'V NP' and 'V *at* NP' frame; besides having the conative interpretation, the *at*-PP may mean "towards" or "in reaction to". Our approach for identifying conative *at*-PPs is not fine-grained enough to make this distinction. This is further aggravated by the sparsity of the conative construction. As a result a larger number of non-alternating verbs was acquired for the conative than for the dative and benefactive alternations (25, 10, and 11 verbs, respectively).

Although a large number of conative verbs was discovered, all verbs were strongly biased towards the transitive frame (see the typicality values in Tables 3.23 and 3.24). This strongly indicates that the conative alternation is not highly frequent. In comparison, neither the dative nor the benefactive alternation show a strong preference for one frame over the other (see the typicality values in Table 3.14). Also note that the productivity estimates should be interpreted together with the typicality ones. On the basis of the productivity values alone, the conative alternation seems fairly productive for SWAT verbs (see Table 3.23). Only by looking at the typicality estimates for this class it becomes apparent that most SWAT verbs are unlikely to alternate (see Tables 3.23 and 3.24). Classes with high productivity values can be interpreted as classes which best exemplify the alternation. For example, GIVE, FUTURE HAVING, MESSAGE TRANSFER and BRING-TAKE verbs are the prototypical example classes for the dative alternation (see Table 3.14), GET verbs for the benefactive alternation (see Table 3.14), and CHEW and SWAT verbs for the conative alternation (see Table 3.23).

Experiment 3 concentrates on the possessor object alternation. Intuitively, we would expect this alternation to be less frequent than the alternations we have studied so far. If this is the case, then we would expect not only low typicality values but also low productivity values for this alternation.

3.4. Experiment 3: The Possessor Object Alternation

3.4.1. Introduction

The possessor object alternation is characterized by a change in the realization of the arguments within the VP. The alternation involves a possessor and a possessed attribute and is illustrated in (3.82) and (3.9), repeated here as (3.83).

(3.82) a. They praised the volunteers' dedication.
b. They praised the volunteers for their dedication. (Levin 1993: 73)
c. They praised the volunteers for your dedication.

(3.83) a. I admired his honesty.
b. I admired him for his honesty. (Levin 1993: 192)
c. I admired him for her honesty.

(3.84) a. I sensed his eagerness.
b. *I sensed him for his eagerness. (Levin 1993: 74)

The alternation concerns the realization of the possessor attribute which can be manifested either as the verbal object (see the nouns *dedication* and *honesty* in (3.82a) and (3.83a), respectively) or as the object of a prepositional phrase headed by *for* (see the PPs *for their dedication* and *for his honesty* in (3.82b) and (3.83b), respectively). In the first case the possessor is expressed as a determiner of the verbal object, whereas in the latter the possessor is

Table 3.25: Sample of verbs with frames characteristic of the possessor object alternation

Alternating	Non-alternating
admire	detect
applaud	discern
hate	feel
value	hear
punish	notice
study	see
support	sense
reproach	smell
want	taste

the direct object. Also note that in the prepositional variant the noun that is the object of the preposition has a possessive determiner which must agree in number, person, and gender with the possessor: examples (3.82c) and (3.83c) cannot be the alternants of (3.82b) and (3.83b), respectively.

Levin (1993) lists 126 which undergo the alternation and 9 verbs for which the prepositional variant is not possible, i.e., these verbs do not allow a possessor object (see example (3.84)). A sample of alternating and non-alternating verbs taken from Levin is given in Table 3.25.

In the following section we describe and evaluate the procedures we used to acquire verbs that license the possessor object alternation. We present our results in Section 3.4.3. Section 3.4.3.1 discusses the productivity and typicality estimates for the possessor object alternation.

3.4.2. Method

3.4.2.1. Acquisition

The experimental setup was the same as in Experiments 1 and 2. The possessor object alternation involves two surface syntactic structures: the transitive 'V NP_{poss}' structure and the prepositional 'V NP1 *for* $NP2_{poss}$' structure. Tokens for the transitive frame 'V NP' were acquired in Experiment 2. From these tokens verbs whose objects contained possessive determiners (i.e., a genitive possessor) were extracted (see the examples in (3.85) below). We obtained 29,245 tokens with possessor objects.

(3.85) a. The Executive Board intended to [$_V$ support] [$_{NP}$ Russia's transition to a market economy].
b. The doctor did not [$_V$ support] [$_{NP}$ the plaintiff's alleged injuries].
c. He [$_V$ supports] [$_{NP}$ the former captain's views].
d. It may [$_V$ support] [$_{NP}$ the family's continued existence].
e. The British Roads Federation [$_V$ supports] [$_{NP}$ the government's new toll road scheme].

Tokens for the prepositional frame 'V NP1 *for* NP2' were extracted in Experiment 1. These tokens were further analyzed in order to separate benefactive from non-benefactive PPs (see Section 3.2.2.6). The prepositional frame poses a different problem for the possessor object alternation: the *for*-PP is causal rather than benefactive and its head noun is modified by a possessive pronoun which agrees in number, gender, and person with the verbal object. In order to consider verbs in the 'V NP1 *for* NP2$_{poss}$' frame as candidates for the possessor object alternation the following requirements need to be met:

1. the prepositional phrase must be attached to the verb;
2. the NP expressing the possessor attribute must have a possessive determiner (i.e., a possessive pronoun);
3. the possessor must agree in number, gender, and person with the possessive determiner.
4. the *for*-PP must be causal.

We addressed requirement 1 in Section 3.2.2.5, where we showed that a simple approach that classifies all instances containing *for*-PPs as verb attachment achieves a precision of 73.9%. Requirement 2 was easily met by extracting tokens with *for*-PPs whose head nouns had possessive determiners (e.g., *my, your, his, her, its, our, their*). We addressed requirement 3 by imposing restrictions on the agreement of the possessive determiner with the possessor. Using a methodology similar to Experiments 1 and 2 we applied the following heuristics.

1. **Accept** if the possessor is a pronoun and agrees in number and gender with the possessor attribute (e.g., *reward them for their vote, remember her for her television series*).
2. **Reject** if the possessor is a pronoun and disagrees in number and gender with the possessor attribute (e.g., *wear it for her lover, bless him for their deliverance*).
3. **Reject** if the verb has been previously attested in the benefactive prepositional frame (e.g., *prepare, make, save*)
4. **Reject** if the possessor is a noun and disagrees in number with the possessor attribute (e.g., *pay the full amount for their drug, use the park for their meetings*).

5. **Cannot decide** if the possessor is a noun and agrees in number with the possessor attribute (e.g., *blame the place for her misfortune*, *love the seashore for its mix*).

Heuristic 5 identified 1,103 tokens. These were dealt with separately by taking into account the gender of the possessor and the possessor attribute. For each possessor NP the head noun was extracted and manually annotated with gender information. Tokens were accepted as instances of the prepositional frame in case of gender agreement between the possessor and the possessor attribute (e.g., *blame the committee for its loss*, *hate the catholic church for its condemnation*). This agreement constraint together with heuristic 1 identified 1,127 tokens.

Requirement 4 concerns the meaning of the preposition *for*. Consider the sentences in (3.86): even though the agreement constraints are satisfied, the *for*-PPs are not characteristic of the possessor object alternation. In (3.86a) the *for*-PP receives a goal interpretation, whereas in (3.86b) it is the argument of the verb *mistake*. We randomly sampled 200 tokens from the 1,127 instances which were identified as having possessor objects in agreement with their possessor attributes. 97.5% of the *for*-PPs contained in the sample were causal ($K = .80$, $N = 200$, $k = 2$). This means that requirement 4 is satisfied via satisfaction of requirements 1–3: in most cases when the *for*-PP is attached to the verb and the possessor agrees in number, gender, and person with the possessive determiner, the *for*-PP is causal.

(3.86) a. Sally and I got ready to take her for her usual walk.
b. The attendant mistook him for his rival.

3.4.2.2. Evaluation

The procedures employed to obtain tokens characteristic of the possessor object alternation were evaluated as follows: 200 tokens were randomly sampled from the 29,245 tokens which were identified as instances of the 'V NP_{poss}' frame. Similarly, the procedure developed for identifying instances of the prepositional frame 'V NP1 *for* $NP2_{poss}$' was evaluated by randomly selecting 200 tokens which were accepted as instances of the frame in question and accordingly 200 tokens which were rejected. Two judges decided whether the tokens were classified correctly.

Not surprisingly tokens characteristic of the V NP_{poss} frame were identified with a high accuracy of 92.4%. The judges reached an agreement of $K = .89$ ($N = 200$, $k = 2$) (see Table 3.26). Recall from Experiment 2 that we achieved an accuracy of 89.3% in detecting transitive tokens in general. Also note that here we only evaluate the correctness of transitive tokens containing objects with possessive determiners (true positives). In Experiment 2 (see Section 3.3.2.4) we evaluated how accurately the proposed methodology identifies tokens which are not instances of transitive verbs (false positives).

The heuristic approach performed well (reaching an average precision of 94.8%) in detecting tokens which are not characteristic of the prepositional frame 'V NP1 *for* $NP2_{poss}$'

Table 3.26: Precision of the heuristics for the possessor object alternation

	Precision	K
Accept V NP_{poss}	92.4%	.89
Accept V NP *for*	82.3%	.78
Reject V NP1 *for*	94.8%	.85

(see heuristics 2–4). The judges' agreement was $K = .85$ ($N = 200$, $k = 2$). The approach also performed well in identifying tokens which are characteristic of the prepositional frame 'V NP1 *for* $NP2_{poss}$' (see Table 3.26). The judges' agreement was $K = .78$ ($N = 200$, $k = 2$).

3.4.2.3. Filtering

In order to eliminate verbs associated with the wrong frame, we discarded verbs with frame frequency less than two. In contrast to Experiments 1 and 2 we did not employ a relative frequency threshold. It was thought that such a threshold would eliminate most verb candidates for the 'V NP1 *for* $NP2_{poss}$' frame for which only a small number of tokens was acquired (1,103 in total). The simple frame frequency cutoff yielded 1,533 verbs with the 'V NP_{poss}' frame and 52 verbs for the 'V NP1 *for* $NP2_{poss}$' frame.

3.4.3. Results

A verb was classified as undergoing the possessor object alternation if it had been found in the corpus with both the 'V NP_{poss}' and 'V NP1 *for* $NP2_{poss}$' frame. According to this criterion 37 verbs were found that license the possessor object alternation. Of these verbs, 22 undergo the alternation and were found both in Levin (1993) and the corpus. From the 15 verbs attested in the corpus only four verbs license the alternation, whereas the remaining 11 do not alternate even though they license both alternating frames. Table 3.27 shows a comparison of the acquired verbs and the verbs found in Levin: rows 'V NP_{poss}' and 'V NP1 *for* $NP2_{poss}$' contain verbs which Levin classifies as alternating but for which only one frame was acquired. Table 3.28 lists the verbs found in the corpus only.

In comparison to Levin's (1993) list of verbs, a small number of verbs were found to alternate in the corpus. This is not surprising given how few verbs were found with the 'V NP1 *for* $NP2_{poss}$' frame (52 in total). Although, the acquired frames were generally correct (see Table 3.28 where no verbs with erroneous frames are listed), approximately one third of the acquired verbs do not license the alternation (see Table 3.28). This points to the limitations of the heuristic approach put forward here which exploits shallow syntactic and grammatical information. Consider the sentences in (3.87)–(3.88): the verb *win* occurs both with the 'V NP_{poss}' and 'V NP1 *for* $NP2_{poss}$' frames. However, without knowing that *women* cannot be the possessor of *race* (see (3.87b)) and correspondingly that *design* is not a possessor attribute of

Table 3.27: Possessor object verbs common in corpus and Levin

Possessor Object Alternation	
Alternating	admire, analyse, commend, condemn, criticise, denounce, despise, enjoy, forgive, hate, honour, like, love, mock, praise, prosecute, punish, repay, respect, reward, value, want
V NP_{poss}	abhor, acclaim, adore, applaud, appreciate, assess, celebrate, censure, compliment, congratulate, denigrate, deplore, detest, dislike, distrust, envy, favour, evaluate, excuse, extol, fancy, fear, fine, greet, hail, insult, lament, miss, mourn, need, penalise, prize, regret, relish, resent, review, ridicule, salute, scorn, scrutinise, shame, snub, study, support, toast, tolerate, trust, welcome, worship
V NP1 *for* $NP2_{poss}$	bless, castigate, chide, compensate, rebuke, reprimand, reproach, scold, upbraid, chastise

Table 3.28: Possessor object verbs found in corpus only

Possessor Object Alternation	
Alternating	attack, curse, fight, remember
Non-alternating	award, blame, help, include, kick, marry, reimburse, seek, slap, win, thank

Civic Trust Award (see (3.88b)) we cannot avoid the conclusion that *win* licenses the possessor object alternation. Another example is the verb *thank* which according to Levin undergoes the alternation. Inspection of the corpus tokens convinced us of the contrary (see the examples in (3.89)–(3.90)).

(3.87) a. Glynis won the women's race.
b. *Glynis won the race for their women.

(3.88) a. This beautiful and unique theatre won the Civic Trust Award for its design.
b. *This beautiful and unique theatre won the design's Civic Trust Award.

(3.89) a. They would like to thank the hotel for their help.
b. *They would like to thank the hotel's help.

(3.90) a. She thanked the Queen for her great kindness.
b. *She thanked the Queen's great kindness.

Levin (1993) defines four semantic classes of verbs which license the possessor object alternation: ADMIRE verbs (e.g., *love*, *dislike*), JUDGMENT verbs (e.g., *criticise*, *applaud*), WANT verbs (e.g., *fancy*, *need*), and ASSESSMENT verbs (e.g., *review*, *scrutinise*). Figure 3.4 compares the acquired verbs against Levin's list of verbs with respect to these four semantic classes. The comparison excluded verbs appearing in Levin with corpus frequency less than

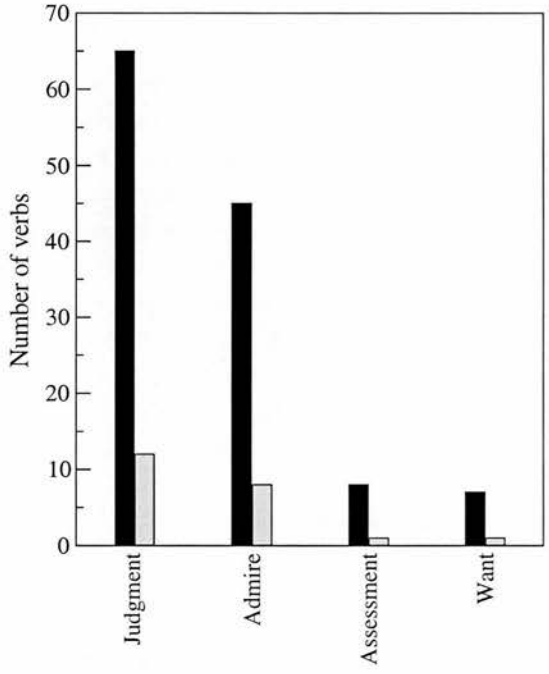

Figure 3.4: Semantic classes for the possessor object alternation

one per million.[14] Although JUDGMENT verbs are the biggest semantic class licensing the alternation both in Levin and the corpus, the number of acquired verbs is small in comparison to Levin (approximately one fifth of the verbs listed in Levin were attested with both frame alternants).

Let us consider now the verbs that license the alternation but are not listed in Levin (1993) (see Table 3.28). Given that ASSESSMENT and WANT verbs are poorly represented in the corpus one would expect the alternation to be more productive for JUDGMENT and ADMIRE verbs. The prediction is borne out: *attack* (when attested with the "criticize" sense), *fight*, and *curse* are JUDGMENT verbs, whereas *remember* is an ADMIRE verb (see the examples in (3.91)–(3.93)).

(3.91) a. He harshly attacked the government 's decision to enter into talks with the Arab states.
b. Mr Kinnock attacked the Government for its policies on housing.

(3.92) a. He cursed the older man's stupidity in insisting Casey be released from the hospital.
b. I curse them for their crass stupidity.

(3.93) a. She remembered Nails's dogged determination.
b. I shall remember him for his fearlessness.

[14]These verbs are *backbite, chasten, decry, defame, deprecate, disdain, disparage, eulogize, execrate, felicitate, idolize, impeach, lambaste, laud, malign, recompense, remunerate, reprove, revile, rue, venerate*, and *vilify*.

Table 3.29: Productivity and typicality estimates for the possessor object alternation

Class	Total	Alt	Prod	AvTyp	StdDev	Min	Max
JUDGMENT	65	12	.18	.44	.28	.12	.99
ADMIRE	45	8	.17	.27	.19	.05	.58
ASSESSMENT	7	1	.14	.12			
WANT	6	1	.16	.14			

Table 3.30: Typicality estimates for the 'V NP1 *for* $NP2_{poss}$ frame

JUDGMENT verbs	Typ	ADMIRE verbs	Typ
commend	.66	admire	.18
condemn	.14	despise	.58
criticize	.18	enjoy	.12
denounce	.12	hate	.44
forgive	.69	like	.09
honour	.19	love	.38
mock	.23	respect	.04
praise	.35	value	.33
prosecute	.50		
punish	.60		
repay	.30		
reward	.73		

3.4.3.1. Productivity and Typicality

We can use the information acquired from the corpus to quantify how productive the possessor object alternation is for a given semantic class. By taking into account the corpus frame frequencies we can estimate how typical the alternation is for a given verb and verb semantic class. The productivity (Prod) and average typicality (AvTyp) values for the possessor object alternation are summarized in Table 3.29. The typicality values have been estimated for the 'V NP1 *for* $NP2_{poss}$' frame (since we report average typicality the standard deviation of the mean (StdDev) and its minimum (Min) and maximum (Max) values are also given in Table 3.29). Table 3.30 shows the individual typicality values for JUDGMENT and ADMIRE verbs.

Not surprisingly the alternation is fairly unproductive for all four semantic classes that license it. JUDGMENT verbs tend to be fairly typical, i.e., they tend to equally prefer the two frames characteristic of the alternation (see Table 3.29), whereas ADMIRE verbs prefer the transitive frame. Since only one verb was acquired for ASSESSMENT and WANT verbs (*analyse* and *want*, respectively) we only report their individual typicality values.

3.4.4. Discussion

Experiment 3 examined the possessor object alternation. This alternation further put to the test our heuristic approach. So far our acquisition procedures took into account syntactic restrictions (e.g., the attachment site of the PPs in the prepositional variants of the dative and benefactive alternations, the detection of compounds which could be mistaken for the double object construction) or semantic restrictions (e.g., the identification of benefactive and conative PPs for the respective alternations). The possessor object alternation not only displays both types of restriction (i.e., the *for*-PP must be causal and also attached to the verb) but also introduces a new type, a grammatical restriction: in the prepositional variant of the alternation the verb object has to agree in number, person, and gender with the PP headed by the preposition *for*. The proposed approach performs well considering the linguistic complexity of diathesis alternations, by taking into account domain-independent shallow syntactic, semantic, and grammatical information.

Experiment 3 further demonstrated, contra Levin (1993), that the possessor object alternation is relatively unattested in the BNC. This is indicated by its low productivity and typicality values (see Tables 3.29 and 3.30). With Experiment 3 we arrive at an interesting characterization of alternations in terms of their productivity and typicality values: for verb classes with high typicality and productivity values a large number of verbs alternates, i.e., they equally prefer either alternating frame (see GIVE and THROWING verbs in Table 3.14). Verb classes with low typicality and productivity values contain verbs that are neither prototypical for the alternation nor are likely to alternate (see ASSESSMENT and WANT verbs in Table 3.29). Verb classes with high productivity and low typicality values, although prototypical for the alternation, are unlikely to alternate (see CHEW verbs in Table 3.23). Finally, verb classes with low productivity and high typicality values are not prototypical for the alternation, although the verbs which undergo the alternation prefer either alternating frame (see PERFORMANCE verbs in Table 3.14).

3.5. General Discussion

The work reported in this chapter relies on type and token frequencies acquired from corpora using partial parsing methods and linguistic heuristics. Frame frequency data acquired through surface cueing was used to estimate the productivity of a semantic class and the typicality of its members. Experiments 1–3 demonstrate how frequency data can be used to quantify linguistic theory: first, our results can be used to empirically investigate whether an alternation is attested or not. Although the alternations studied in this chapter are well represented in Levin (1993) in terms of the number of verbs that license them, our experiments revealed that only dative and benefactive verbs are sufficiently attested in the corpus in both frame alternants. Second, corpus data can be used to complement manual linguistic classifications. Note that throughout

Table 3.31: Overall productivity values for four alternations

Alternation	Total	Alt	OvProd
Dative	115	45	.40
Benefactive	103	33	.32
Conative	79	36	.46
Possessor Object	126	22	.26

our experiments we found alternating verbs which are not listed in Levin. These verbs can be further used to expand and validate Levin's semantic classification. For example, comparison between novel verbs licensing the alternation and verbs listed in Levin can assess whether novel verbs are compatible with Levin's semantic classes or even reveal additional semantic classes undergoing the alternation (see Experiment 2).

Let us consider more closely the results of Experiments 1–3. We can quantify the overall productivity of an alternation (OvProd) as the overall number of verbs that were found to undergo the alternation in the corpus over the number of verbs that license the alternation and are listed in Levin (1993). The overall productivity values for the four alternations studied in Experiments 1–3 are shown in Table 3.31. Under the assumption that the number of acquired alternating verbs indicates how frequent the alternation is, we would expect, if we were to look for these four alternations in a different corpus, to find more dative and benefactive than possessor object verbs. The conative verbs are particularly interesting: approximately half of the verbs listed in Levin were found in the corpus; this result alone indicates that the conative alternation is fairly productive. However, none of the acquired verbs are typical for the alternation: the typicality values range from .06 to .18 (see Table 3.23) suggesting that conative verbs are biased towards the transitive frame. Given these typicality values, we would expect to find a small number of alternating conative verbs, especially in a corpus which is smaller than the BNC.

Consider now the productivity values which are estimated relative to a verb semantic class. If the productivity values are indeed indicative of how easily the alternation applies for a given semantic class, then we would expect verbs of productive classes to be attested in corpora more often than verbs of unproductive classes. For example, for the dative alternation we would expect to acquire more verbs from the fairly productive classes of FUTURE HAVING, GIVE, MESSAGE TRANSFER, and BRING-TAKE verbs than from the unproductive classes of SLIDE, DRIVE, and INSTRUMENT OF COMMUNICATION verbs (see Table 3.14). Similarly, for the benefactive alternation we would expect to find a larger number of GET verbs than CREATE or BUILD verbs (see Table 3.14).

The productivity and typicality measures not only capture the empirical properties of alternating verbs but also enable us to formulate explicit predictions about their behavior. We test these predictions in Experiment 4 by looking at the Penn Treebank corpus (Marcus

et al. 1994, see Chapter 2 for details). This corpus differs from the BNC in three important aspects: (a) it is considerably smaller (approximately one million words), (b) it exhibits a uniform writing style (it is a collection of newspaper articles), and (c) it is annotated not only with part-of-speech but also with phrase structure information. In the first part of Experiment 4 we describe how frames characteristic of the dative, benefactive, conative, and possessor object alternation were extracted from the Penn Treebank corpus. The second part attempts to answer the following questions: (a) are alternations frequent/infrequent across corpora and (b) how predictive are the BNC class productivity estimates of the types of verbs we expect to find in a different corpus?

3.6. Experiment 4: Validation

3.6.1. Introduction

In this experiment we further investigate the usefulness of the productivity and typicality measures. We test the empirical validity of these measures by determining the extent to which their predictions hold invariably across corpora. In Experiment 4 we acquire frames characteristic of the dative, benefactive, conative, and possessor object alternation from a domain specific corpus, the Penn Treebank (a collection of financial texts taken from the Wall Street Journal, Marcus et al. 1994) and compare our results against the BNC. On the basis of the results of Experiments 1–3, Experiment 4 tests the following predictions:

1. Some alternations are more frequent than others. A diagnostic for measuring the frequency of the alternation is the number of verbs for which the alternation is acquired (see the overall productivity estimates in Table 3.31) in conjunction with the typicality values of its semantic classes. Assuming there are no corpus genre effects, we predict that more verbs will be acquired for the dative and benefactive than for the conative and possessor object alternation. The former alternations are not only relatively frequent (see the overall productivity values Table 3.31) but also are licensed by several fairly typical classes (see Table 3.14). The possessor object alternation is neither productive (see Table 3.31) nor is licensed by classes which are typical with respect to the alternating frames (see Table 3.29). The conative alternation is also licensed by atypical classes (see Table 3.23), even though a large number of conative verbs are attested in the BNC (see Table 3.31).

2. Verbs of certain semantic classes are more likely to alternate than others. This means that we expect to acquire more verbs for productive classes. Based on the results of Experiment 1, we predict that BRING-TAKE, FUTURE HAVING, GIVE, and MESSAGE TRANSFER verbs are more likely to undergo the dative alternation than CARRY, DRIVE, or SLIDE verbs (see Table 3.14). Similarly, GET verbs are more likely to

license the benefactive alternation than BUILD or PERFORMANCE verbs (see Table 3.14). Based on the results of Experiment 2 we predict that SWAT and CHEW verbs are the most likely to undergo the conative alternation. Both classes are well represented in the corpus (the productivity value of CHEW and SWAT is .80, see Table 3.23) and display relatively high typicality values when compared for instance to HIT or CUT verbs (.18 for CHEW and .12 for Swat, see Table 3.23). Finally, the results of Experiment 3 indicate that none of the four verb semantic classes that license the possessor object alternation are particularly productive or typical (see Table 3.29).

In the following we describe how we extracted frames characteristic of the four alternations from the Penn Treebank and present and discuss our results.

3.6.2. Method

Frames characteristic of the dative, benefactive, conative, and possessor object alternation were extracted using *tgrep*, a tool which permits the search of the Penn Treebank for syntactic patterns based on a user-specified query. The tool uses regular expressions and a query language (developed for the manipulation of tree structured data) to search for syntactic structures of arbitrary complexity. Recall from Chapter 2 that the Penn Treebank is annotated with syntactic information and therefore the frame extraction process can be fairly accurate: arguments and adjuncts are given distinct labels; for arguments several grammatical functions are distinguished (e.g., subject, object, predicate), and adjuncts are further labeled with the semantic roles of direction, location, manner, purpose, and time; discontinuous constituents, coordination, and gapping are also annotated.

We used *tgrep* to search for the syntactic patterns in (3.94)–(3.96). Pattern (3.94a) identified ditransitive verbs (600 tokens). Prepositional arguments are distinguished from prepositional adjuncts in the Penn Treebank and therefore pattern (3.94b) only identified verbs which genuinely subcategorize for an NP and a *to*-PP (353 tokens). Pattern (3.94b) only looked for adjunct *for*-PPs. Tokens matching pattern (3.94b) (524 in total) were manually inspected so as to distinguish benefactive from non-benefactive *for*-PPs. The benefactive tokens were 162. Pattern (3.95a) identified transitive verbs (23,447 tokens), and pattern (3.95b) identified adjunct *at*-PPs. Since adjunct PPs are labeled with their semantic roles, the pattern only matched directive *at*-PPs (5 tokens). Pattern (3.96a) matched transitive verbs with possessive objects (59 tokens), whereas pattern (3.96b) identified tokens whose verbal objects agreed in number and person with the NP headed by the preposition *for* (54 tokens). These tokens were manually inspected in order to distinguish causal *for*-PPs (19 tokens).

(3.94) a. V NP1 NP2 gives the artist a sense of purpose
b. V NP1 *to* NP2 issue cards to the public
c. V NP1 *for* NP2 win military aid for the rebels

(3.95) a. V NP hit the door
b. V *at* NP pick at the food

(3.96) a. V NP_{poss} praised the department's actions
b. V NP1 *for* $NP2_{poss}$ love my son for his character

The extracted tokens were further lemmatized using Abney's (1997) stemmer. The process yielded 56 verbs with the 'V NP1 NP2' frame, 40 verbs with the 'V NP1 *to* NP2' frame, 89 verbs with the 'V NP1 *for* NP2' frame, 2,105 transitive verbs, four verbs with the 'V *at* NP' frame, 28 verbs with the 'V NP_{poss}' frame and 13 verbs with the 'V NP1 *for* $NP2_{poss}$' frame. The acquired verb frames were not submitted to additional filtering (compare Experiments 1–3), since the extraction process relied on the Treebank parses and grammatical distinctions which were thought to be fairly accurate.

3.6.3. Results

As in Experiments 1–3 a verb was classified as undergoing an alternation if it had been found in the corpus with both frames characteristic of the alternation. According to this criterion 21 verbs were found that license the dative alternation. Of these verbs, 16 undergo the dative alternation and are listed in Levin (1993) (see Table 3.32), two license the alternation but are not listed in Levin (*hand* and *deliver*) and three license the two frames without licensing the alternation (e.g., *get*, *provide*). 13 verbs were found to license the benefactive alternation: of these verbs, six undergo the alternation and were found both in the corpus and Levin (see Table 3.32), two undergo the alternation but were not included in Levin (*do* and *provide*), and five do not alternate even though they license both frames (e.g., *take*, *sell*, *offer*). One verb was found to license the conative alternation which was also listed in Levin and finally no verbs were found that license the possessor object alternation.

Table 3.32 shows the common verbs between Levin (1993) and the corpus for the dative, benefactive, and conative alternation (recall that no verbs were acquired for the possessor object alternation). Most verbs were acquired for the dative alternation. A few verbs were acquired for the benefactive alternation, whereas the conative and possessor object alternation are not well represented in the Treebank corpus. Table 3.32 also shows the distribution of semantic classes for the acquired dative, benefactive, and conative verbs (only the verbs common between Levin and the corpus are displayed). Levin lists 10 semantic classes which undergo the dative alternation: five of these were found in the corpus; most acquired verbs belong to FUTURE HAVING and GIVE verbs (six and five, respectively). There are five classes which license the benefactive alternation in Levin. Of these classes only the class of GET verbs was found in the corpus. Of the eight classes that license the conative alternation in Levin only CHEW verbs were attested in the Penn Treebank.

Table 3.32: Alternating verbs common in the Penn Treebank and Levin

Dative Alternation	
FUTURE HAVING	award, extend, grant, offer, owe, promise
GIVE	give, lend, loan, pay, sell
M. TRANSFER	show, tell
SEND	mail send
BRING-TAKE	bring
Benefactive Alternation	
GET	buy, earn, gain, leave, save, win
Conative Alternation	
CHEW	pick

3.6.4. Discussion

The results of Experiment 4 confirm our initial predictions: more verbs were acquired for alternations with high overall productivity and class typicality values. More specifically, more dative and benefactive verbs were acquired in comparison to conative or possessor object verbs (compare Tables 3.31 and 3.32).

Another important result concerns the acquired verbs and their semantic classes. Recall from Experiment 1 that the most productive classes for the dative alternation were BRING-TAKE, FUTURE HAVING, GIVE, MESSAGE TRANSFER, and SEND verbs with productivity values 1.00, .73, .73, .70, and .44, respectively (see Table 3.14). The dative verbs found in the Penn Treebank belong precisely to these five classes (see Table 3.32). In Experiment 2 we found that GET verbs were the most productive class for the benefactive alternation with a productivity value of .51 (see Table 3.14). Only GET verbs were found in the Penn Treebank. Finally, CHEW and SWAT verbs were found to be the most productive classes for the conative alternation with productivity value of .8 (see Table 3.23). The single verb we acquired from the Penn Treebank is a member of the class of CHEW verbs (see Table 3.32). Also note that CHEW verbs are the most typical class for the conative alternation (its average typicality value is .18, see Table 3.23).

Obviously we do not expect a perfect match between the two corpora. It has been shown that corpus idiosyncrasies can affect subcategorization frequencies. Roland and Jurafsky (2000) extracted verb frame frequencies from a number of different corpora and demonstrated that corpus type (written versus spoken) and discourse type (single sentences versus connected discourse) influence the verb frame frequencies. This suggests that different corpora may give different results with respect to verb alternations. The number of acquired verbs and their frame frequencies expectedly varies from the BNC to the Penn Treebank. The two corpora vary in size (the Penn Treebank is considerably smaller than the BNC, one million words versus 100 million) and corpus register (domain specific versus wide coverage text). Furthermore, the

acquisition method employed in Experiment 4 relied on the syntactic annotation available with the Penn Treebank; a combination of shallow parsing and linguistic heuristics was adopted for the acquisition studies in Experiments 1–3.

Describing alternations in terms of their productivity and typicality enables us to observe their behavior across corpora. Instead of focusing on individual verbs and their observed differences, we can investigate the empirical properties of different semantic classes and different alternations. Comparison of the Penn Treebank with the BNC revealed that our productivity and typicality measures do not simply express statistical tendencies in a particular corpus using a particular methodology (i.e., shallow parsing) but rather hold across corpora and methods.

3.7. Related Work

The acquisition of alternations from corpora combines work on the induction of subcategorization frames and the extraction of lexical semantic information. There has been a considerable amount of work on the acquisition of subcategorization frames from corpora primarily aiming at the automatic construction of subcategorization dictionaries (Brent 1993; Briscoe and Carroll 1997; Manning 1993). In some cases the acquired frames have been used for studies relating to diathesis alternations (McCarthy 2000; Schulte im Walde 2000). Levin's (1993) seminal study on diathesis alternations and verb semantic classes has recently influenced work in dictionary creation (Dang, Rosenzweig, and Palmer 1997), machine translation (Dorr 1997), generation (Stede 1998), and automatic lexical acquisition (Korhonen 1998; McCarthy 2000; McCarthy and Korhonen 1998; Merlo and Stevenson 1999; Resnik 1993; Schulte im Walde 1998, 2000; Stevenson and Merlo 2000). In this section we discuss work relating to the induction of subcategorization information from corpora and the acquisition of diathesis alternations. Approaches that either make use of or directly try to reproduce Levin's semantic classification are discussed in the following chapter.

Brent's (1993) work on frame acquisition uses morphosyntactic cues to extract subcategorization information from an unannotated corpus. A given word is considered a verb if and only if it occurs in the corpus both with and without the suffix *-ing*. Once potential verbs are found, their argument structure is identified by taking surface syntactic cues into account. For example, a verb is considered transitive if it is followed by an accusative pronoun and punctuation or by an accusative pronoun and a temporal conjunction such as *when*. Brent acquires only six subcategorization frames ('V NP', 'V THAT', 'V INF', 'V NP THAT', 'V NP INF', and 'V NP NP'). Verbs with erroneous frames are discarded by applying hypothesis testing on binomial frequency data (see Section 3.2.2.7 for details). Brent's procedure achieves a considerably high precision (i.e., proportion of subcategorization frames that the method acquires correctly) ranging from 96.0% to 100%, whereas recall (i.e., number subcategorization frames learned over the number of frames that exist for a given verb) is somewhat lower ranging from 47.0%

to 100%.

Brent's (1993) subcategorization acquisition method relies on highly reliable syntactic cues (only cues containing pronouns as indicators of subjects and objects are used) and consequently does not scale up to a greater number of subcategorization frames for which reliable syntactic cues may not be available. The shortcomings of Brent's approach are addressed by Manning (1993) who first uses a stochastic part-of-speech tagger (Kupiec 1992) to annotate the corpus and then a finite-state parser to detect verbs and their complements. Erroneous subcategorization frames are discarded using Brent's method of hypothesis testing. Manning acquires 19 different subcategorization frames and evaluates the results of his method on 40 randomly selected verbs against the subcategorizations listed in the Oxford Advanced Learner's Dictionary of Current English (OALD, Hornby 1989). Manning's approach achieves a precision of 90.0% and a recall of 43%. Token recall reaches approximately 82.0% (for 200 verbs) when measured as the number of verbs which are accounted for by the acquired subcategorization dictionary in a random sample of text (selected from the corpus on which the subcategorization experiment was performed).

Briscoe and Carroll (1997) extend Manning's (1993) work by acquiring a larger number of subcategorization frames (160 in total). Briscoe and Carroll use a statistical parser which yields complete though shallow parses (e.g., no analysis of unbounded dependencies is attempted). Their method consists of the following steps: raw text is part-of-speech tagged (Elworthy 1994) and lemmatized; the annotated text is parsed by a probabilistic LR parser using a unification-based grammar; a patternset extractor extracts subcategorization patterns from the ranked parses, and a pattern classifier assigns the extracted patterns to 160 subcategorization frames. Erroneous subcategorization frames are filtered by a modified version of Brent's (1993) original hypothesis testing. The system is evaluated on 14 verbs against a subcategorization dictionary which was created by merging the entries of the Alvey Natural Language Tools dictionary (ANLT, Boguraev, Briscoe, Carroll, Carter, and Grover 1987) and the COMLEX dictionary (Grishman et al. 1994) and also against manual analysis of the corpus data (7 verbs). The system achieves a precision of 65.7% and 76.6% against the dictionary and the corpus, respectively. Recall is 35.5% and 43.4% against the dictionary and the corpus, respectively. The system's token recall is 80.9%. A comparison between Manning's and Briscoe and Carroll's results is not possible since the two methods use corpora which differ in size, a different number of subcategorization frames, and are evaluated on different sets of verbs.

Schulte im Walde (1998) extracts subcategorization frames from the BNC using a robust statistical parser (Carroll and Rooth 1998). The parser utilizes a probabilistic context-free grammar (PCFG) for English. In a PCFG, context-free rules are annotated with probabilities, and the probability of a parse is computed as the product of the probabilities of the relevant rules. The grammar is an extension of the standard PCFG model in that it incorporates information about the lexical heads of constituents. Such a head-lexicalized PCFG provides a

grammar model that combines lexicalized rules and lexical coherence relations between constituents. The parameters of this model were iteratively trained on a tagged version of the BNC by applying the Expectation Maximization algorithm (EM, Dempster, Laird, and Rubin 1977). Schulte im Walde acquires 88 different subcategorization frames. Erroneous frames are not discarded through hypothesis testing. Instead a relative frame frequency threshold is applied (i.e., only frames with frequency larger than 5.0% of the total verb frequency are considered as valid frames for a given verb). The acquired frames are not evaluated against an existing subcategorization dictionary but are further used to cluster verbs according to their subcategorization behavior with the aim of discovering verb clusters which bear similarities to Levin's (1993) distinctions of verb semantic classes (see the discussion in the following chapter).

We acquired frames characteristic of the dative, benefactive, conative, and possessor object alternation using a methodology similar to Manning's (1993). Verbs and their arguments were extracted from a part-of-speech annotated and lemmatized version of the BNC using Gsearch as a shallow parser. The approach differs from Briscoe and Carroll (1997) and Schulte im Walde (1998) in that no complete analysis of corpus sentences was performed. Furthermore, the present study narrowly focused on the frames characteristic for the dative, benefactive, conative, and possessor object alternation. Our methodology also bears similarities to Brent's (1993) use of surface syntactic cues. In particular, linguistic heuristics were used so as to identify relatively correct frame tokens (see Sections 3.2.2.2, 3.3.2.2, and 3.4.2.1). Our approach not only takes surface syntactic structure into account but also employs semantic cues in order to acquire frame tokens with meanings corresponding to the alternations at hand. In order to acquire frames for the conative and benefactive alternations we distinguished benefactive *for*-PPs from non-benefactive ones and conative *at*-PPs from non-conative ones (see Sections 3.2.2.6 and 3.3.2.3).

Although we did not use a statistical parser as Briscoe and Carroll (1997) and Schulte im Walde (1998), we used probabilistic information (i.e., Hindle and Rooth's 1993 LA-score based on the log-likelihood ratio) in deciding the attachment site of ambiguous PPs (see Section 3.2.2.4). The log-likelihood ratio was also used to detect compound nouns which could be mistaken for the double object construction (see Section 3.2.2.2). Erroneous frames were not discarded through hypothesis testing, a method which favors subcategorization frames that are well-attested (Manning and Schütze 1999: 276). It was thought that hypothesis testing would penalize verbs with correct but low frequency frames. Instead, we opted for a simpler approach which takes relative frame frequency into account (see Section 3.2.2.7).

The acquired frames were not evaluated against an existing subcategorization dictionary since the focus of Experiments 1–4 was the acquisition of alternating verbs only and not of a general purpose subcategorization dictionary. Subcategorization dictionaries do not consistently list verb alternations. They are built by lexicographers for human readers and are prone to errors, inconsistencies and omissions (Briscoe and Carroll 1997). Furthermore, our results

indicate that alternations are often novel and even Levin's (1993) resource does not provide a full list of verbs and their alternation behavior. The acquired alternating verbs were compared against Levin and the verbs which were not listed in Levin were manually inspected. For the dative alternation 75 verbs were acquired with both frame variants: 80.0% of these verbs had the acquired frames. For the benefactive alternation 72.9% of the alternating verbs (70 in total) had the acquired frames. Of the 101 verbs which were found to license the conative alternation, 83.2% had the acquired frames. Finally, of the 37 verbs that were found to license the possessor object alternation, 100% had the acquired frames.

The main objective of the work presented in the previous sections (one aspect of which is the acquisition of subcategorization frames) was to examine the extent to which Levin's (1993) generalizations about diathesis alternations are attested in a large corpus. We have shown that corpus frequencies can be used to quantify linguistic intuitions by using the acquired frame frequencies to estimate the productivity of an alternation for a given semantic class and the typicality of its members. It has been argued that knowledge about diathesis alternations can be useful for a variety of tasks in NLP.

Korhonen (1998) argues that knowledge about the alternating behavior of verbs can improve the automatic acquisition of subcategorization dictionaries. For example, if two frames have been acquired for a given verb and we know that the two frames do not constitute a valid alternation, then we should discard the frame for which we have less evidence. Stede (1998) shows how knowledge about verb alternations can be exploited within the context of generation, in particular since subtle changes in meaning may correspond to distinct surface syntactic realizations. Dorr (1997) argues that Levin's (1993) semantic classification of verbs into groups exhibiting similar alternation behavior can be useful for the development of dictionaries with rich semantic representations for foreign language tutoring and machine translation.

Work on the acquisition of alternations from corpora has combined information about verb subcategorization and selectional preferences. Resnik (1993) uses his model of selectional restrictions (which is based on the information theoretic notion of relative entropy) to identify verbs participating in the unspecified object alternation (e.g., *Mike ate the cake* versus *Mike ate*). Resnik tests his model on 15 alternating and 19 non-alternating verbs and demonstrates that verbs that license the alternation select more strongly for direct objects than verbs that do not. Resnik's approach is somewhat difficult to evaluate, since there is no precision or baseline reported.

McCarthy and Korhonen (1998) attempt to automatically *identify* verbs participating in diathesis alternations. They extract subcategorization frames from corpus data using Briscoe and Carroll's (1997) acquisition algorithm. Selectional preferences for the acquired frames are obtained using a method initially proposed by Li and Abe (1995) which exploits the WordNet hierarchy. The acquired frames are compared against Levin's (1993) index of alternating verbs and a mapping is defined between the acquired frames and each alternation pair given in Levin.

McCarthy and Korhonen use the causative-inchoative alternation as a test case. They choose 15 verbs taking the frames involved in this alternation and show that the Minimum Description Length Principle (MDL) is good at detecting whether a verb participates in the alternation or not. The system achieves an accuracy of 87.0% when compared against a human judge.

McCarthy (2000) acquires subcategorization frames and selectional preferences as in McCarthy and Korhonen (1998). However, McCarthy abandons MDL in favor of a measure of distributional similarity which compares the similarity of selectional preferences of a given verb under the assumption that alternating verbs exhibit more similar selectional preferences than non-alternating ones. The approach is tested on the causative and conative alternation: 46 verbs are chosen that license the causative alternation and 46 verbs that do not; similarly, 6 alternating and 6 non-alternating verbs are chosen for the conative alternation. The method reaches an accuracy of 73.0% for the causative alternation and an accuracy of 67.0% for the conative alternation, when the argument slots over which the distributional similarity is computed are WordNet classes. A lower precision of 60.0% and 58.0%, for the causative and conative alternation, respectively, is achieved when the argument slots are lemmas instead of WordNet classes.

McCarthy's (2000) approach applies only to what she calls *role switching alternations*. These are alternations where a given argument may have different grammatical realizations (e.g., subject or object) while maintaining the same thematic role (see example (3.1) in Section 3.1 where the theme *the cup* is the object in (3.1a) and the subject in (3.1b)). McCarthy claims that the approach does not require apriori knowledge specific to the alternation i.e., the methodology is the same irrespectively of the syntactic particularities of the alternation under investigation as long as it qualifies as role switching. However, the method heavily relies on WordNet for the acquisition of selectional restrictions and also on knowledge of the subcategorization frames involved in a given alternation.

Our approach is considerably more shallow than McCarthy and Korhonen's (1998) and McCarthy's (2000). We do not perform full-scale acquisition of subcategorization frames and do not acquire selectional preferences. We do not attempt to classify a given verb as alternating or not: verbs with frames characteristic of the alternation at hand are manually distinguished into verbs that genuinely alternate and verbs that do not alternate but have the alternating frames. McCarthy's approach could be used to automatically distinguish the verbs we acquire into alternating and non-alternating. Despite the fact that our approach is shallow and does not make use of semantic knowledge in the form of selectional restrictions, the number of non-alternating verbs acquired through the procedures described in the previous sections is relatively small. 75.0% of the acquired dative verbs (i.e., verbs with frames characteristic of the dative alternation) license the dative alternation, 78.4% of the acquired benefactive verbs license the benefactive alternation, 70.2% of the acquired conative verbs license the conative alternation, and 70.3% of the possessor object verbs license the possessor object alternation.

A common feature in all previous work is that Levin's (1993) semantic classification and index of alternations are used together with other knowledge sources (i.e., WordNet) or automatic learning techniques (i.e., acquisition of subcategorization frames and selectional restrictions) in order to aid lexicographic work, parsing, or machine translation and therefore it is implicitly assumed that Levin's generalizations have an empirical basis. The work presented here tests precisely this assumption. We examine whether actual corpus data provides evidence for diathesis alternations by exploiting surface syntactic and semantic cues. We use corpus data to examine the extent to which different alternations are productive and also to detect a given alternation's representative members and show that, even a shallow approach, can be useful for quantifying and complementing linguistic generalizations.

Our approach makes use of surface syntactic, semantic, and grammatical cues. Although no selectional restrictions are used, we identify a large number of alternating verbs with relatively high accuracy. The acquired frame token frequencies can be used to indicate how likely it is for a given verb or verb class to alternate. The acquired frame type frequencies can be used to indicate the "fit" between the alternation and the semantic classes for which it applies. For example, GET verbs are the core semantic class for the benefactive alternation (see Table 3.14), which is intuitively correct. Replication of our results using a different corpus (see Experiment 4 in Section 3.6) confirms that our observations are not the result of the statistical behavior in a particular corpus.

3.8. Summary

In this chapter we assessed the empirical validity of Levin's (1993) theory of diathesis alternations. Experiments 1–3 examined the extent to which diathesis alternations are attested in corpus data using shallow syntactic and semantic processing. Alternating verbs were acquired from the BNC by using Gsearch as a chunk parser. Erroneous frames were discarded by applying linguistic heuristics, statistical scores (the log-likelihood ratio) and large-scale lexical resources (e.g., WordNet).

We argued that productivity and typicality can be used to characterize the alternating behavior of individual verbs and their semantic classes and to make explicit predictions about word use. Experiment 4 tested the validity of these measures by determining the extent to which they give results that hold across corpora.

We showed that corpus frequencies can be used to quantify and complement linguistic generalizations using a shallow approach which looks for surface syntactic rather than deep semantic regularities. In the following chapter we make use of the acquired frequencies in a probabilistic framework which focuses on semantically ambiguous verbs. More specifically, we concentrate on verbs which receive multiple classifications in Levin (1993) and demonstrate how the proposed approach, which relies on surface cues and approximations, can be combined

with a probabilistic model so as to constrain the verb class ambiguity inherent in Levin by placing a preference ordering on the space of possible verb meanings.

Chapter 4

A Probabilistic Model of Verb Class Ambiguity

In Chapter 3 we examined how the surface cueing approach can be used for the acquisition of alternating verbs and their respective frames. We also proposed *productivity* and *typicality* as descriptors of the empirical behavior of alternating verbs. We showed that (a) alternations differ in terms of their overall productivity, (a) semantic classes differ in terms of their likelihood to alternate, and (c) verbs exhibit preferences over frame alternants. Furthermore, these tendencies are observed across corpora.

In this chapter we exploit the acquired frequencies in a probabilistic framework that constrains the verb class ambiguity arising from Levin's (1993) taxonomy of verbs and their classes, a widely used resource for lexical semantics. In her framework, some verbs, such as *give* exhibit no class ambiguity. But other verbs, such as *write*, can be a member of several distinct classes. In some of these ambiguous cases, the appropriate class for a particular token of a verb is immediately obvious from inspection of the surrounding context. In others it is not, and an application which wants to recover this information will be forced to rely on some more or less elaborate process of inference. In this chapter we present a simple statistical model of verb class ambiguity and show how it can be used to carry out such inference. We recast Levin's classification in a probabilistic framework (using corpus-based distributions) and show that corpora provide a rich resource for testing and quantifying linguistic theory.

4.1. Introduction

The relation between the syntactic realization of a verb's arguments and its meaning has been extensively studied in Levin (1993) (see Chapter 3). Levin's work relies on the hypothesis that the behavior of a verb, particularly with respect to the expression and interpretation of its arguments, is to a large extent determined by its meaning (Levin 1993: 1). Verbs which display

the same *diathesis alternations*—alternations in the realization of their argument structure—are assumed to share certain meaning components and are organized into a semantically coherent class.

A methodological consequence of this assumption is that verb behavior (i.e., participation in diathesis alternations) can be used to provide clues about aspects of meaning, which in turn can be exploited to characterize verb senses (classes in Levin's 1993 terminology). These verb senses are defined at a higher level of granularity than traditionally assumed in lexicographic work concerned with sense distinctions (e.g., WordNet or machine-readable dictionaries). As an example consider examples (3.7), (3.8), (3.47), and (3.83) repeated here as (4.1), (4.2), (4.4), and (4.6). Examples (4.1) and (4.2) illustrate the dative and benefactive alternations. The conative and possessor object alternation are illustrated in (4.4) and (4.6), respectively (see Chapter 3 for details).

(4.1) a. Bill sold a car to Tom.
b. Bill sold Tom a car.

(4.2) a. Martha carved the baby a toy.
b. Martha carved a toy for the baby.

(4.3) a. Martha carved the toy out of the piece of wood.
b. Martha carved the piece of wood into a toy. (Levin 1993: 173)

(4.4) a. Paula hit the fence.
b. Paula hit at the fence.

(4.5) a. Paula hit the fence with the stick.
b. The stick hit the fence. (Levin 1993: 149)

(4.6) a. I admired his honesty.
b. I admired him for his honesty.

Observation of the semantic and syntactic behavior of *pay* and *give* reveals that they pattern with *sell* in licensing the dative alternation (see example (4.1)). These verbs are all members of the GIVE class. Verbs like *make* and *build* pattern with *carve*; they license not only the benefactive alternation (see (4.2)) but also the material/product alternation (see example (4.3)). These verbs are representative of the class of BUILD verbs. Verbs *beat* and *kick* behave similarly to *hit* in undergoing the conative and instrument subject alternation (see examples (4.4)) and (4.5), respectively). These verbs are all members of the HIT class. Finally, the verb *admire* together with the semantically related *adore*, *cherish*, and *enjoy* undergo the possessor object alternation (see (4.6)). By grouping together verbs which pattern together with respect to diathesis alternations, Levin (1993) defines approximately 200 verb classes, which she argues reflect important semantic regularities.

Levin's (1993) taxonomy of verbs explores the systematic correspondence between meaning and argument structure. As a result, subcategorization patterns can give clues about the

Table 4.1: Polysemous verbs according to Levin

Classes	Verbs	BNC frequency
1	2,239	4,252,715
2	536	2,325,982
3	173	738,854
4	43	395,212
5	23	222,747
6	7	272,669
7	2	26,123
10	1	4,427

meaning of a verb and vice versa knowledge of the semantics of a given verb can be predictive of the realization of its arguments. For example, members of the HIT class (e.g., *hit, kick*) inherit information about the alternations they license (e.g., conative, with/against, instrument subject alternation, etc.) and verbs undergoing the possessor object alternation will predictably be JUDGMENT, ADMIRE, ASSESSMENT, or WANT verbs (see Section 3.4 for details).

Levin (1993) provides an index of 3,024 verbs for which she lists the semantic classes and diathesis alternations. The mapping between verbs and classes is not one-to-one. Of the 3,024 verbs which she covers, 784 are listed as having more than one class. Even though Levin's monosemous verbs outnumber her polysemous verbs by a factor of nearly four to one, the total frequency of the former (4,252,715) is comparable to the total frequency of the latter (3,986,014). This means that close to half of the cases processed by a hypothetical semantic tagger would manifest some degree of ambiguity. The frequencies are detailed in Table 4.1 and were compiled from a lemmatized version of BNC. Furthermore, as shown in Figure 4.1, the number of alternations licensed by a given verb increases with the number of classes it inhabits. Consider for example verbs participating in one alternation only: of these, 90.4% have one semantic class, 8.6% have two classes, .7% have three classes, and .3% have four classes. In contrast, of the verbs licensing six different alternations, 14% have one class, 17% have two classes, 12.4% have three classes, 53.6% have four classes, 2% have six classes, and 1% has seven classes.

Palmer (2000) and Dang, Kipper, Palmer, and Rosenzweig (1998) argue that the use of syntactic frames and verb classes can simplify the definition of different verb senses. Beyond this, we argue that information about the argument structure of a polysemous verb can often help disambiguate it. Consider for instance the verb *serve*, which is a member of four classes: GIVE, FIT, MASQUERADE, and FULFILLING. Each of these classes can in turn license four distinct syntactic frames. As shown in the examples below, in (4.7a) *serve* appears ditransitively and belongs to the semantic class of GIVE verbs, in (4.7b) it occurs transitively and is a member of the class of FIT verbs, in (4.7c) it takes the predicative complement *as minister of the interior* and is a member of MASQUERADE verbs. Finally, in sentence (4.7d) *serve* is a FULFILLING

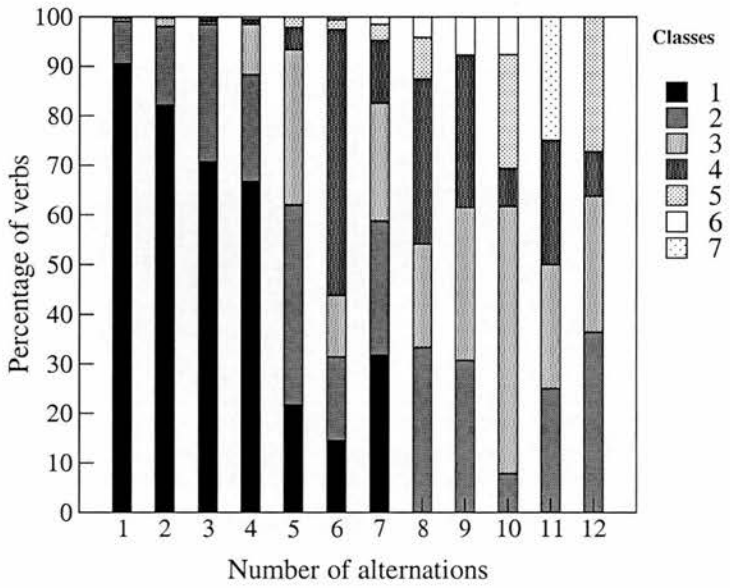

Figure 4.1: Relation between number of classes and alternations

verb and takes two complements, a noun phrase (*an apprenticeship*) and a prepositional phrase headed by *to* (*to a still-life photographer*). In the case of verbs like *serve* we can guess their semantic class solely on the basis of the frame with which they appear.

(4.7) a. I'm desperately trying to find a venue for the reception which can serve our guests an authentic Italian meal.

b. The airline serves 164 destinations in over 75 countries.

c. Jean-Antoine Chaptal was a brilliant chemist and technocrat who served Napoleon as minister of the interior from 1800 to 1805.

d. Before her brief exposure to pop stardom, she served an apprenticeship to a still-life photographer.

But sometimes we do not have the syntactic information that would provide cues for semantic disambiguation. Consider example (4.8). The verb *write* is a member of three Levin classes, two of which (MESSAGE TRANSFER, PERFORMANCE) take the double object frame. In this case, we have the choice between the MESSAGE TRANSFER reading (see (4.8a)) and the PERFORMANCE reading (see (4.8b)). The same situation arises with the verb *toast* which is listed as a PREPARE verb and a JUDGMENT verb; both these classes license the prepositional frame 'NP1 V NP2 *for* NP3'. In sentence (4.9a) the preferred reading is that of PREPARE instead of JUDGMENT (see sentence (4.9b)). The verb *study* is ambiguous among three classes when attested in the transitive frame: LEARN (see example (4.10a)), SIGHT (see example (4.10b)), and ASSESSMENT (see example (4.10c)). The verb *convey* when attested in the prepositional frame 'NP1 V NP2 *to* NP3' can be ambiguous between the SAY class (see example (4.11a)) and the SEND class (see example (4.11b)). In order to infer the semantic class

for a given ambiguous verb, we would not only need detailed semantic information about the verb's arguments but also a considerable amount of world knowledge. For example, in (4.8) selectional restrictions alone do not provide enough clues to disambiguate (4.8a) from (4.8b) since both *letter* and *screenplay* are written material. However, world knowledge indicates that although both scripts and letters can be written for someone, only letters can be written to someone.

(4.8) a. A solicitor wrote him a letter at the airport.
b. I want you to write me a screenplay called "The Trip".

(4.9) a. He sat by the fire and toasted a piece of bread for himself.
b. We all toasted Nigel for his recovery.

(4.10) a. Chapman studied medicine at Cambridge.
b. Romanov studied the old man carefully, looking for some sign that he knew exactly what had been awaiting him at the bank.
c. The alliance will also study the possibility of providing service to other high-volume products, such as IBM and multi-vendor workstations.

(4.11) a. By conveying the news to her sister, she would convey by implication something of her own anxiety.
b. The judge signed the committal warrant and the police conveyed Mr. Butler to prison, giving the warrant to the governor.

The phenomenon is widespread among a variety of classes and frames (e.g., double object, transitive, prepositional frame, see examples (4.8)–(4.11)). The ambiguity exhibited by Levin's (1993) classification is by no means surprising. Verbal meaning is specified without exhaustively enumerating the similarities among semantically related verbs by isolating linguistically pertinent aspects of meaning and using them to minimize the amount of information necessary to specify the syntactic and semantic properties for any given verb. While exploiting systematic correspondences between verbal argument structure and meaning, Levin's objective is to represent all possible meanings for a given verb rather than to provide information about the likelihood of those meanings.

The ambiguity exhibited by verbs like *write* or *study* (see examples (4.8) and (4.10)) is an instance of the common problem of inferring the value of a hidden variable (in this case the "true class" of a particular instance of *write* or *study*). In this chapter, we address the verb class disambiguation problem by developing a probabilistic framework which combines linguistic knowledge (i.e., Levin's 1993 classification) and frame frequencies acquired from the BNC. More specifically, we exploit corpus-derived frequencies in a probabilistic model which places a preference ordering on the set of possible meanings (i.e., semantic classes) without taking discourse or pragmatic information into account. We determine for a given verb and its frame its most likely meaning overall (i.e., across the corpus) instead of focusing on the meaning of individual corpus tokens (see examples (4.8)–(4.11)). The dominant meaning in the absence of

explicit contextual information is modeled probabilistically in a Bayesian framework in which distributional information (in the form of conditional probabilities) is combined with linguistic generalizations (in the form of prior probabilities). Our approach relies on simple surface cues and approximations based on Levin's classification without taking selectional restrictions or world knowledge into account. We are thus able to explore the extent to which an unsupervised learning approach which relies on gross simplifications about the correspondences of surface structure and meaning yields satisfactory results.

Our experiments focus on polysemous verbs with frames characteristic of the dative, benefactive, conative, and possessor object alternations (see Chapter 3). These frames are licensed by a fairly large number of classes and are therefore likely to exhibit class ambiguity: 20 classes license the double object frame, 22 license the prepositional frame 'NP1 V NP2 *to* NP3', 17 classes license the benefactive 'NP1 V NP2 *for* NP3' frame, 118 (out of 200) classes license the transitive frame, and 15 classes license the conative 'NP1 V *at* NP2' frame. In Section 4.2 we describe the probabilistic model and the estimation of the various model parameters. In Experiments 5 and 6 we use the model to derive the dominant class for polysemous verbs and evaluate its performance.

4.2. The Model

Consider again the sentences in (4.8). Assuming that we more often write something to someone rather than for someone, we would like to derive MESSAGE TRANSFER as the prevalent class for *write* rather than PERFORMANCE. We view the choice of a class for a polysemous verb in a given frame as the joint probability $P(c,f,v)$ where v is the polysemous verb subcategorizing for the frame f and inhabiting more than one class c. By choosing the ordering $\langle v,f,c\rangle$ for the variables c, f, and v we can rewrite $P(c,f,v)$ using the chain rule as follows.

$$P(c,f,v) = P(v) \cdot P(f|v) \cdot P(c|v,f) \tag{4.12}$$

Although the parameters $P(v)$ and $P(f|v)$ can be estimated from the BNC ($P(v)$ reduces to the number of times a verb is attested in the corpus and $P(f|v)$ can be obtained through parsing), the estimation of $P(c|v,f)$ is somewhat problematic since it relies on the frequency $f(c,v,f)$ as shown below:

$$P(c|v,f) = \frac{f(c,v,f)}{f(v,f)} \tag{4.13}$$

It would be straightforward to estimate $f(c,v,f)$ if we had access to a parsed corpus annotated with subcategorization and semantic class information. Lacking such as corpus we will assume that the semantic class determines the subcategorization patterns of its members

independently of their identity (see (4.14)).

$$P(c|v,f) \approx P(c|f) \tag{4.14}$$

The independence assumption is a simplification of Levin's hypothesis that the argument structure of a given verb is a direct reflection of its meaning. Moreover, we assume that verbs of the same class uniformly subcategorize (or not) for a given frame. This is evidently not true for all classes of verbs. For example, all GIVE verbs undergo the dative diathesis alternation and therefore we would expect them to be attested in both the double object and prepositional frame, but only a subset of CREATE verbs undergo the benefactive alternation. For example, the verb *invent* is a CREATE verb and can be attested only in the benefactive prepositional frame (*I will invent a tool for you* versus ?*I will invent you a tool*, see Levin 1993 for details). By applying Bayes Law we write $P(c|f)$ as:

$$P(c|f) = \frac{P(f|c) \cdot P(c)}{P(f)} \tag{4.15}$$

By substituting (4.15) into (4.12), $P(c,f,v)$ can be written as:

$$P(c,f,v) \approx \frac{P(v) \cdot P(f|v) \cdot P(f|c) \cdot P(c)}{P(f)} \tag{4.16}$$

We estimate the probabilities $P(v)$, $P(f|v)$, $P(f|c)$, and $P(c)$ as follows:

$$P(v) = \frac{f(v)}{\sum_i f(v_i)} \tag{4.17}$$

$$P(f|v) = \frac{f(f,v)}{f(v)} \tag{4.18}$$

$$P(f|c) = \frac{f(f,c)}{f(c)} \tag{4.19}$$

$$P(c) = \frac{f(c)}{\sum_i f(c_i)} \tag{4.20}$$

$$P(f) = \frac{f(f)}{\sum_i f(f_i)} \tag{4.21}$$

Table 4.2: Sample of verb classes and their syntactic frames

Class	Frame
WIPE MANNER	NP1 V NP2 *from* NP3, NP1 V NP2, NP1 V *at* NP2, NP1 V NP2 AP
ACCOMPANY	NP1 V NP2, NP1 V NP2 *to* NP3
THROW	NP1 V NP2 NP3, NP1 V NP2 *loc* NP3, NP1 V NP2 *from* NP3 *to* NP4, NP1 V NP2, NP1 V NP2 *to* NP3, NP1 V NP2 *at* NP3
PERFORMANCE	NP1 V, NP1 V NP2, NP1 V NP2 NP3, NP1 V NP2 *to* NP3, NP1 V NP2 *for* NP3, NP1 V NP2
GIVE	NP1 V NP2 *to* NP3, NP1 V NP2 NP3
CONTRIBUTE	NP1 V NP2 *to* NP3

It is easy to obtain $f(v)$ from the lemmatized BNC. We automatically acquired syntactic frames for the dative, benefactive, conative, and possessor object alternation using Gsearch (Corley et al. 2001), a tool which facilitates search of arbitrary part-of-speech tagged corpora for shallow syntactic patterns based on a user-specified context-free grammar and a syntactic query (see Chapter 2 for details). The acquisition and filtering process is detailed in Chapter 3. We rely on Gsearch to provide moderately accurate information about verb frames in the same way that Hindle and Rooth (1993) relied on Fidditch to provide moderately accurate information about syntactic structure, and Ratnaparkhi (1998) relied on simple heuristics defined over part-of-speech tags to deliver information nearly as useful as that provided by Fidditch. We estimated $f(f,v)$ as the number of times a verb co-occurred with a particular frame in the corpus.

We cannot read off $P(f|c)$ directly from the corpus, because it is not annotated with verb classes. Nevertheless we can use the information listed in Levin (1993) with respect to the syntactic frames exhibited by the verbs of a given class. Consider PERFORMANCE verbs for example. They are intransitive (e.g., *Mary sang*), transitive (e.g., *Mary sang a song*), and ditransitive when attested in the double object (e.g., *Mary sang a song to me*, *Mary sang me a song*) or benefactive frame (e.g., *Mary sang a song for me*). For each class we recorded the syntactic frames it licenses (see Table 4.2). Levin's (1993) description of the argument structure of various verbs goes beyond the simple listing of their subcategorization. Useful information is provided about the thematic roles of verbal arguments and their interpretation. Consider the examples in (3.11) repeated here as (4.22): in (4.22a) the verb *present* is a member of the FULFILLING class and its theme is expressed by the prepositional phrase *with an award*, in (4.22b) the PP headed by *with* receives a locative interpretation and the verb *load* inhabits the SPRAY/LOAD class, whereas in (4.22c) the prepositional phrase is instrumental and *hit* inhabits the HIT class. None of the information concerning thematic roles was retained. All three classes (FULFILLING, SPRAY/LOAD, and HIT) were assigned the frame 'NP1 V *with* NP2'.

(4.22) a. John presented the student with an award.

b. John loaded the truck with bricks.

c. John hit the wall with a hammer.

Because we did not have corpus counts for the quantity $f(f,c)$ we simply assumed that all frames for a given class are equally likely. This means, for instance, that the estimate for P(NP1 V NP2 *to* NP3|GIVE) is $\frac{1}{2}$ (since the GIVE classes licenses two frames, i.e., 'NP1 V NP2 *to* NP3' and 'NP1 V NP2 NP3') and similarly the estimate for P(NP V|PERFORMANCE) is $\frac{1}{6}$ (see Table 4.2). This is clearly a simplification, since one would expect $f(f,c)$ to be different for different corpora, and to vary with respect to the number of allowable alternations. Furthermore, we have shown in Chapter 3 that not all classes are equally typical with respect to an alternation (i.e., the likelihood of a class attested with a certain frame varies across classes and frames). In order to estimate $P(c)$, we first estimate $f(c)$ which we rewrite as follows:

$$f(c) = \sum_i f(v_i, c) \tag{4.23}$$

Note that for monosemous verbs the estimate of $f(v,c)$ reduces to the count of the verb in the corpus. Once again we cannot estimate $f(v,c)$ for polysemous verbs directly. The task would be straightforward if we had a corpus of verbs, each labeled explicitly with class information. All we have is the overall frequency of a given verb in the BNC and the number of classes it is a member of according to Levin (1993). Since polysemous verbs can generally be the realization of more than one semantic class, counts of semantic classes can be constructed by dividing the contribution from the verb by the number of classes it belongs to (Lauer 1995; Resnik 1993):

$$f(v,c) \approx \frac{f(v)}{|classes(v)|} \tag{4.24}$$

Here, $f(v)$ is the number of times the verb v was observed in the corpus and $|classes(v)|$ is the number of classes c it belongs to. For example, in order to estimate the frequency of the class GIVE we consider all verbs that are listed as members of this class in Levin's (1993) taxonomy. The class contains 13 verbs, among which six are polysemous. We will estimate f(GIVE) by taking into account the verb frequency of the monosemous verbs ($|classes(v)|$ is one in this case) as well as distributing the frequency of the polysemous verbs among their classes. For example, *feed* inhabits the classes GIVE, GORGE, FEEDING, and FIT and occurs in the corpus 3,263 times. We will increment the count of f(GIVE) by $\frac{3,263}{4}$. Table 4.3 illustrates the estimation of $f(v,c)$ for all members of the GIVE class. The total frequency of the class is obtained by summing over individual the $f(v,c)$'s (see equation (4.23)).

Note that the estimation in (4.24) relies on the simplifying assumption that the frequency of a verb is distributed evenly across its semantic classes. This is clearly not true for all verbs. Consider for example the verb *rent* which inhabits classes GIVE (*Frank rented Peter his*

Table 4.3: Estimation of $f(\text{GIVE})$

v	$\lvert classes(v)\rvert$	$f(v)$	$f(v, \text{GIVE})$
give	1	126,894	126,894
lend	1	2,650	2,650
loan	1	198	198
pay	1	34,794	34,794
peddle	1	140	140
refund	1	184	184
sell	1	19,904	19,904
trade	2	2,570	1285
serve	4	15,457	3,864.25
lease	2	524	262
rent	2	1,060	530
feed	4	3,263	815.75
repay	2	1,089	544.5

room) and GET (*I rented my flat for my sister*). Intuitively speaking, the GIVE sense of *rent* is more frequent than GET, however this is not taken into account by the estimation scheme in (4.24), primarily because we do not know the true distribution of the classes for *rent*. A more informed estimation scheme would distribute the verb frequency unequally among verb classes by taking into account the class size. For instance, if we knew that GIVE verbs are more likely than GET verbs, we could take this fact into account while distributing the frequency of *rent* between these two classes by assigning the GIVE class a larger proportion of the verb's frequency. Informally this means that bigger classes are given more weight than smaller classes. We can derive an estimate of class size by taking into account the number of verbs inhabiting it. In what follows we formally describe this alternative estimation scheme for $P(c)$ (see equations (4.20) and (4.23)).

We rewrite the frequency $f(v,c)$ as in (4.25). Note that we cannot estimate $P(c|v)$, the true distribution of the verb among its classes, directly from the corpus. We approximate $P(c|v)$ by collapsing across all verbs that have the appropriate pattern of ambiguity (see (4.26)):

$$f(v,c) = f(v) \cdot P(c|v) \tag{4.25}$$

$$f(v,c) \approx f(v) \cdot P(c|amb\_class) \tag{4.26}$$

Here *amb_class*, the ambiguity class of a verb, is the set of classes that it might inhabit.[1] We collapse verbs into ambiguity classes in order to reduce the number of parameters which must

[1] Our use of ambiguity classes is inspired by a similar use in HMM based part-of-speech tagging (Kupiec 1992).

Table 4.4: Estimation of $f(v,c)$ for the verb *feed*

c	$size(c)$	$P(c\|amb\_class)$	$f(v,c)$
GIVE	15	.39	1,272.57
GORGE	8	.21	685.23
FEED	3	.08	261.04
FIT	12	.32	1,044.16

be estimated: we certainly lose information, but the approximation makes it easier to get reliable estimates from limited data. We simply approximate $P(c|amb\_class)$ using a heuristic based on class size:

$$p(c|amb\_class) \approx \frac{size(c)}{\sum_{c \in amb\_class} size(c)} \tag{4.27}$$

For each class we recorded the number of its members after discarding verbs whose frequency was less than one per million in the BNC. This gave us a first approximation of the size of each class. We then computed, for each polysemous verb, the total size of the classes of which it was a member. We calculated $P(c|amb\_class)$ by dividing the former by the latter (see equation (4.27)). We obtained an estimate for the class frequency $f(c)$ by multiplying $P(c|amb\_class)$ by the observed frequency of the verb in the BNC (see equation (4.26)).

As an example consider again the estimation of $f(\text{GIVE})$. In order to estimate the contribution of the verb *feed* we need to distribute its corpus frequency among the classes GIVE, GORGE, FEED, FIT. The respective $P(c|amb\_class)$ for these classes are $\frac{15}{38}$, $\frac{8}{38}$, $\frac{3}{38}$, and $\frac{12}{38}$. By multiplying these by the frequency of *feed* in the BNC (3,263) we obtain the estimates for $f(v,c)$ given in Table 4.4. Only the frequency $f(feed, \text{GIVE})$ is relevant for the estimation of $f(\text{GIVE})$. The estimation process described above involves at least one gross simplification, since $P(c|amb\_class)$ is calculated without reference to the identity of the verb in question. For any two verbs which fall into the same set of classes $P(c|amb\_class)$ will be the same, even though one or both may be atypical in its distribution across the classes. Furthermore, the estimation tends to favor large classes, again irrespectively of the identity of the verb in question. For example, the verb *carry* has three classes, CARRY, FIT, and COST. Intuitively speaking, the CARRY class is the most frequent (e.g., *Smoking can impair the blood which carries oxygen to the brain*, *I carry sugar lumps around with me*). However, since the FIT class (e.g., *Thameslink presently carries 20,000 passengers daily*) is larger than the CARRY class, it will be given a higher probability (.45 versus .4). This is clearly wrong, but it is an empirical question how much it matters.

Tables 4.5 and 4.6 show the ten most frequent classes as estimated using (4.24) and (4.26). We explore the contribution of the two estimation schemes for $f(c)$ in Experiments 5 and 6. Note that simply relying on class size, without regard to verb frequency, would

Table 4.5: The ten most frequent classes using equal distribution of verb frequencies

c	$f(c)$
CHARACTERIZE	601,647.4
GET	514,308.0
SAY	450,444.6
CONJECTURE	390,618.4
FUTURE HAVING	369,229.3
DECLARE	264,923.6
AMUSE	258,857.9
DIRECTED MOTION	252,775.6
MESSAGE TRANSFER	248,238.7
GIVE	208,884.1

Table 4.6: The ten most frequent classes using unequal distribution of verb frequencies

c	$f(c)$
GET	453,843.6
SAY	447,044.2
CHARACTERIZE	404,734.2
CONJECTURE	382,193.8
FUTURE HAVING	370,717.7
DECLARE	285,431.7
DIRECTED MOTION	255,821.6
POCKET	247,392.7
AMUSE	205,729.4
GIVE	197,828.8

give quite different results for $P(c)$ (see equations (4.26) and (4.27)). For example, the class of MANNER OF SPEAKING verbs has 76 members, of which 30 have frequencies which are less than one per million, and is the seventh largest class in Levin's classification. According to our estimation schemes (see (4.24) and (4.26)) MANNER OF SPEAKING verbs are the 116th largest class.

Finally, we wanted to estimate the probability of a given frame, $P(f)$ (see equation (4.21)). We could have done this by acquiring Levin compatible subcategorization frames from the BNC. Techniques for the automatic acquisition of subcategorization dictionaries have been developed by, among others, Manning (1993), Briscoe and Carroll (1997), and Carroll and Rooth (1998) (see Section 3.7 in Chapter 3 for details). But the present study was less ambitious, and narrowly focused on the frames representing the dative, benefactive, conative, and possessor object alternation. The estimation of $P(f)$ was carried out by taking into account Levin's (1993) distribution of frames. In particular, by counting the number of times a given frame is licensed by several semantic classes we get an estimate of the frequency of the frame $f(f)$. Dividing frame frequency by the total number of frames licensed by all semantic classes gives us the probability distribution $P(f)$ (see equation 4.21). Table 4.7 shows the ten most likely frames as derived by this estimation procedure.

The probabilities $P(f|c)$ and $P(f|v)$ will be unreliable when the frequency estimates for $f(f,v)$ and $f(f,c)$ are small, and ill-defined when the frequency estimates are zero. Following Hindle and Rooth (1993) we smooth the observed frequencies in the following way, where $f(f,V) = \sum_i f(f,v_i)$, $f(V) = \sum_i f(v_i)$, $f(f,C) = \sum_i f(f,c_i)$ and $f(C) = \sum_i f(c_i)$. We redefine

Table 4.7: Ten most likely frames

Frame	$P(f)$
NP1 V NP2	.20
NP1 V	.14
NP1 V NP2 *with* NP3	.05
NP1 V *loc* NP2	.05
NP1 V NP2 AP	.05
NP1 V NP2 *to* NP3	.03
NP1 V NP2 NP3	.03
NP1 V NP2 *loc* NP3	.03
NP1 V NP2 *from* NP3	.03
NP1 V *with* NP2	.02
NP1 V NP2 *for* NP3	.02

the probability estimates as follows:

$$P(f|v) \approx \frac{f(f,v) + \frac{f(f,V)}{f(V)}}{f(v)+1} \tag{4.28}$$

$$P(f|c) \approx \frac{f(f,c) + \frac{f(f,C)}{f(C)}}{f(c)+1} \tag{4.29}$$

When $f(f,v)$ is zero, the estimate used is proportional to the average $\frac{f(f,V)}{f(V)}$ across all verbs. Similarly, when $f(f,c)$ is zero, our estimate is proportional to the average $\frac{f(f,C)}{f(C)}$ across all classes. We do not claim that this scheme is perfect, but any deficiencies it may have are almost certainly masked by the effects of approximations and simplifications elsewhere in the system.

In Experiment 5 we use the model to test the hypothesis that subcategorization information can be used to disambiguate polysemous verbs. In particular, we concentrate on verbs like *serve* (see example (4.7)) which can be disambiguated solely on the basis of their frame. In Experiment 6 we focus on verbs which are genuinely ambiguous, i.e., they inhabit a single frame and yet can be members of more than one semantic class (e.g., *write*, *study*, see examples (4.8)–(4.11) in Section 4.1). In this case, we use the probabilistic model to assign a probability to each class the verb inhabits. The class with the highest probability represents the dominant meaning for a given verb.

4.3. Experiment 5: Using Subcategorization to Resolve Verb Class Ambiguity

4.3.1. Method

We evaluated the performance of the model presented in Section 4.2 on all verbs listed in Levin (1993) which are polysemous and take frames characteristic of the dative, benefactive, conative, and possessor object alternations (see Chapter 3). In this experiment we focused solely on verbs whose meaning can be potentially disambiguated by taking into account their subcategorization frame. We considered 128 verbs with the double object frame (2.72 average class ambiguity), 101 verbs with the prepositional frame 'NP1 V NP2 *to* NP3' (2.59 average class ambiguity), 113 verbs with the frame 'NP1 V NP2 *for* NP3' (2.63 average class ambiguity), 42 verbs with the frame 'NP1 V *at* NP3' (3.05 average class ambiguity), and 39 verbs with the transitive frame (2.28 average class ambiguity).

The task was the following: given that we know the frame of a given verb can we predict its semantic class? In other words by varying the class c in the term $P(c,f,v)$ we are trying to see whether the class which maximizes it is the one predicted by the lexical semantics and the argument structure of the verb in question. The model's responses were evaluated against Levin's (1993) classification. The model's performance was considered correct if it agreed with Levin in assigning a verb to an appropriate class given a particular frame.

Recall from Section 4.2 that we proposed two approaches for the estimation of the class probability $P(c)$. We explore the influence of the parameter $P(c)$ on the model's performance by obtaining two sets of results corresponding to the two estimation schemes.

4.3.2. Results

The model's accuracy is shown in Tables 4.8 and 4.9. The results in Table 4.8 were obtained using the estimation scheme for $P(c)$ which relies on the even distribution of the frequency of a verb across its semantic classes (see equation (4.24)). The results in Table 4.9 were obtained using an alternative scheme which distributes verb frequency unequally among verb classes by taking class size into account (see equation (4.26)). As mentioned in Section 4.3.1, the results were obtained by comparing the model's performance against Levin's (1993) classification. We also compared the results to the baseline of choosing the most frequent class (without taking subcategorization information into account). Recall from Section 4.2 that we obtained estimates for class frequencies $f(c)$ in order to estimate $P(c)$ (see equation (4.20)). Given a class ambiguous verb, our naive baseline procedure defaulted to the class with the highest frequency $f(c)$, where class frequency was determined by the estimation procedures described in Section 4.2 (see equations (4.23), (4.24), (4.26), and (4.27)).

The model achieved a precision of 93.9% using either type of estimation for $P(c)$. It

Table 4.8: Model precision using equal distribution of verb frequencies for the estimation of $P(c)$

Frame	Baseline	Model
NP1 V NP2 NP3	60.9%	93.8%
NP1 V NP *to* NP3	63.3%	95.0%
NP1 V NP *for* NP3	63.6%	98.2%
NP1 V *at* NP2	2.4%	83.3%
NP1 V NP2	43.6%	87.2%
Combined	55.8%	93.9%

Table 4.9: Model precision using unequal distribution of verb frequencies for the estimation of $P(c)$

Frame	Baseline	Model
NP1 V NP2 NP3	62.5%	93.8%
NP1 V NP *to* NP3	67.3%	95.0%
NP1 V NP *for* NP3	66.4%	98.2%
NP1 V *at* NP2	2.4%	85.7%
NP1 V NP2	41.0%	84.6%
Combined	56.7%	93.9%

also outperformed the baseline by 38.1% (see Table 4.8) and 37.2% (see Table 4.9). One might expect a precision of 100% since these verbs can be disambiguated solely on the basis of their frame. However, the performance of our model is less, mainly because of the way we estimated the terms $P(c)$ and $P(f|c)$: we over-emphasize the importance of class information without taking into account how individual verbs distribute across classes. Note that the two estimation schemes yield comparable performances. This is a positive result given the importance of $P(c)$ in the estimation of $P(c,f,v)$. Recall that we are trying to find the class c maximizing the probability $P(c,f,v)$ (see equation (4.16)). Our model simply uses probability distributions acquired from a large corpus (e.g., verb frequencies, frame frequencies) in combination with prior linguistic knowledge (i.e., Levin's 1993 taxonomy). Although our estimation schemes are not perfect, they do not seem to yield counterintuitive results.

4.3.3. Discussion

Our results in the previous section exploit systematic correspondences between meaning and subcategorization in a statistical framework. Comparison against a naive baseline shows that important information is lost when the systematic relation between meaning and syntax is not taken into account.

Despite the simplicity of the model introduced in Section 4.2, the estimation of its parameters assumes information which is not readily available in the corpus (unless it is annotated with verb class information). We approximated several model parameters generally adopting a heuristic approach which combines frequencies derived from the corpus with linguistic information inherent in Levin's (1993) classification. The results in Section 4.3.2 indicate that the approximations yield satisfactory performance.

A more demanding task for our probabilistic model will be with genuinely ambiguous verbs (i.e., verbs for which the mapping between meaning and subcategorization is not one-to-one). Although native speakers may have intuitions about the dominant interpretation for a given verb, this information is entirely absent from Levin (1993). In Experiment 6 we show

how our model can be used to recover this information.

4.4. Experiment 6: Using Corpus Distributions to Derive Verb Class Preferences

4.4.1. Method

We tested the performance of our model on 67 genuinely ambiguous verbs, i.e., verbs which inhabit a single frame and can be members of more than one semantic class (e.g., *write*). These verbs were listed in Levin (1993) and undergo the dative, benefactive, conative, and possessor object alternations (see Chapter 3). As in Experiment 5, we considered verbs with the double object frame (3.27 average class ambiguity), verbs with the frame 'NP1 V NP2 *to* NP3' (2.94 average class ambiguity), verbs with the frame 'NP1 V NP2 *for* NP3' (2.42 average class ambiguity), verbs with the frame 'NP1 V *at* NP3' (2.71 average class ambiguity), and transitive verbs (2.77 average class ambiguity). The model's predictions were compared against manually annotated data. More specifically, corpus tokens characteristic of the verb and frame in question were randomly sampled from the BNC and annotated with class information so as to derive the "true" distribution of the verb's classes in a particular frame. We describe the verb selection procedure as follows.

Given the restriction that these verbs are semantically ambiguous in a specific syntactic frame we could not simply sample from the entire BNC, since this would decrease the chances of finding the verb in the frame we are interested in. Instead, for all class ambiguous verbs tokens were randomly sampled from the parsed data used for the acquisition of frame frequencies for the dative, benefactive, conative, and possessor object alternations (see Chapter 3). The model was evaluated on verbs for which a reliable sample could be obtained. This meant that verbs had to have a frame frequency larger than 50. For verbs exceeding this threshold 100 tokens were randomly selected and annotated with verb class information (see the details below). For verbs with frame frequency less than 100 and more than 50, no sampling took place, the entire set of tokens was manually annotated. This selection procedure resulted in 14 verbs with the double object frame, 16 verbs with the frame 'NP1 V NP2 *to* NP3', two verbs with the frame 'NP1 V NP2 *for* NP3', one verb with the frame 'NP1 V *at* NP3', and 80 verbs with the transitive frame. From the transitive verbs we further randomly selected 34 verbs; these were manually annotated and used for evaluating the model's performance.[2]

The selected tokens were annotated with class information by two judges. The classes were taken from Levin (1993) and augmented with the class OTHER which was reserved for corpus tokens which either had the wrong frame or for which the classes in question were not

[2]Although the model can yield predictions for any number of verbs, evaluation could not be performed for all 80 verbs for which our judges would have to annotate 8,000 corpus tokens.

Table 4.10: Model precision using equal distribution of verb frequencies for the estimation of $P(c)$

Frame	Baseline	Model
NP1 V NP2 NP3	50.0%	78.6%
NP1 V NP *to* NP3	43.8%	68.8%
NP1 V NP *for* NP3	00.0%	100.0%
NP1 V *at* NP2	100.0%	100.0%
NP1 V NP2	47.1%	73.5%
Combined	46.2%	74.6%

Table 4.11: Model precision using unequal distribution of verb frequencies for the estimation of $P(c)$

Frame	Baseline	Model
NP1 V NP2 NP3	50.0%	78.6%
NP1 V NP *to* NP3	43.8%	75.0%
NP1 V NP *for* NP3	00.0%	100.0%
NP1 V *at* NP2	100.0%	100.0%
NP1 V NP2	47.1%	67.6%
Combined	46.2%	73.1%

applicable. The judges were given annotation guidelines (for each verb) but no prior training (the guidelines are given in Appendix A, Section A.1). The annotation not only provided a gold standard for evaluating the model's performance, it also enabled us to examine the empirical basis of Levin's classification and to test whether humans agree on the class annotation task. We measured the judges' agreement on the annotation task using the Kappa coefficient (Cohen 1960, see Section 2.5.1 for details). We counted the performance of our model as correct if it agreed with the "most preferred", i.e., most frequent verb class, as determined in the manually annotated corpus sample by taking the average of the responses of both judges.

As in Experiment 5, we explored the influence of the parameter $P(c)$ on the model's performance by obtaining two sets of results corresponding to the two estimation schemes discussed in Section 4.2.

4.4.2. Results

The model's accuracy is shown in Tables 4.10 and 4.11. The results in Table 4.11 were obtained using the estimation scheme for $P(c)$ which relies on the even distribution of a verb's frequency across its semantic classes (see equation (4.24)). The results in Table 4.10 were obtained using a scheme which distributes verb frequency unequally among verb classes by taking class size into account (see equation (4.26)). As in Experiment 5, the results were compared to a simple baseline which defaults to the most frequent class without taking verb frame information into account (see equations (4.23), (4.24), (4.26), and (4.27) in Section 4.2).

The model achieved a precision of 74.6% using the estimation scheme of equal distribution and a precision of 73.1% using the estimation scheme of unequal distribution. The difference between the two estimation schemes is not statistically significant (using the χ^2 statistic $p = .84, N = 67$). Tables 4.12–4.16 give the distribution of classes for the 67 polysemous verbs as obtained from the manual annotation of corpus tokens together with inter-annotator agreement (K). The presence of the symbol $\surd$ indicates that the model's class preference for a given verb agrees with its distribution in the corpus. The absence of $\surd$ indicates disagreement. For the

Table 4.12: Semantic preferences for verbs with the double object frame

Verb		Class				K
call	√	DUB 93	GET 3	OTHER 4		.82
cook		BUILD 28	PREPARE 33	OTHER 1		1.00
declare	√	DECLARE 35	REF. APPEAR. 18	OTHER 5		.89
feed	√	FEED 61	GIVE 32	OTHER 6		.73
find	√	DECLARE 36	GET 47	OTHER 17		.70
leave	√	GET 6	FULFILL 14	F. HAVE 56	OTHER 23	.67
make	√	BUILD 21	DUB 66	OTHER 13		.79
pass	√	GIVE 81	SEND 0	THROW 0	OTHER 19	.93
save	√	BILL 24	GET 62	OTHER 14		.74
shoot		THROW 91	GET 0	OTHER 5		1.00
take		BRING-TAKE 15	PERFORM 40	OTHER 45		.77
write	√	MSG. TRANS. 54	PERFORM 19	OTHER 18		.85

comparison shown in Tables 4.12–4.16 model class preferences were derived using the equal distribution estimation scheme for $P(c)$ (see equation (4.24)).

As shown in Tables 4.12–4.16[3] the model's predictions are generally borne out in the corpus data. Misclassifications are due mainly to the fact that the model introduced in Section 4.2 does not take verb-class dependencies into account. Consider for example the verb *cook* in Table 4.12. According to the model the most likely class for *cook* is BUILD. Although it may generally be the case that BUILD verbs (e.g., *make*, *assemble*, *build*) are more frequent than PREPARE verbs (e.g., *bake*, *roast*, *boil*) the situation is reversed for *cook*. The same is true for the verb *shoot* which when attested in the double object frame is more likely to be a THROW verb (*Jamie shot Mary a glance*) rather than a GET verb (*I will shoot you two birds*). Notice that our model is not context-sensitive, i.e., it does not derive class rankings tailored to specific verbs, primarily because this information is not readily available in the corpus as explained in Section 4.2. A similar situation manifests itself with the verbs *express* and *pose* in Table 4.13,

[3] MSG. TRANS. abbreviates the class of MESSAGE TRANSFER verbs, REF. APPEAR. abbreviates REFLEXIVE VERBS OF APPEARANCE, F. HAVE abbreviates FUTURE HAVING verbs, PERFORM abbreviates PERFORMANCE verbs, CONTR. abbreviates CONTRIBUTE verbs, CONT. LOC. abbreviates VERBS OF CONTIGUOUS LOCATION, S. EMISSION abbreviates VERBS OF SOUND EMISSION, W. MANNER abbreviates WIPE MANNER VERBS, and BODY-INT. MOTION abbreviates VERBS OF BODY-INTERNAL MOTION.

Table 4.13: Semantic preferences for verbs with the 'NP1 V NP2 *to* NP3' frame

Verb		Class				K
convey	√	SAY 40	SEND 57	OTHER 3		.77
express		SEND 87	REF. APPEAR. 80	OTHER 10		.74
extend		F. HAVE 50	CONTR. 37	OTHER 13		.71
fly	√	DRIVE 35	RUN 24	OTHER 10		.85
issue	√	F. HAVE 52	FULFILL 42	OTHER 5		.68
leave		F. HAVE 28	FULFILL 70	OTHER 2		.72
offer	√	F. HAVE 87	REF. APPEAR. 0	OTHER 12		.66
pass	√	GIVE 60	SEND 26	THROW 5	OTHER 9	.70
pose		MSG. TRANS. 61	REF. APPEAR. 0	OTHER 34		.88
present	√	FULFILL 79	REF. APPEAR. 19	OTHER 2		.94
return	√	SEND 31	CONTR. 56	OTHER 6		.82
show	√	MSG. TRANS. 57	REF. APPEAR. 0	OTHER 39		.84
suggest	√	SAY 62	REF. APPEAR. 8	OTHER 10		.73
serve	√	GIVE 36	FULFILL 12	OTHER 10		.74
take	√	PERFORM 52	CREATE 13	OTHER 33		.77
tell	√	MSG. TRANS 73	TELL 5	OTHER 22		.98
transfer	√	SEND 31	CONTR. 58	OTHER 11		.69

where in both cases the REFLEXIVE VERBS OF APPEARANCE class is preferred over the more intuitive classes SEND and MESSAGE TRANSFER.

The verbs *savour*, *support*, and *value* are respectively assigned the classes SIGHT, CONTIGUOUS LOCATION, and PRICE (see Table 4.16) instead of the more intuitive ADMIRE class, even though the latter is the most frequent of all four classes (SIGHT is 42nd, CONTIGUOUS LOCATION is 23rd, PRICE is 116th, and ADMIRE is 20th). In this case, the estimation of the parameter $P(f|c)$ is harmful. Recall from Section 4.2 that we assumed that all frames are equally likely for a given class. This means that the probability $P(f|c)$ is larger for classes with a small number of frames. The class ADMIRE licenses five frames in contrast to SIGHT, CON-

Table 4.14: Semantic preferences for verbs with 'NP1 V NP2 *for* NP' and 'NP1 V *at* NP2' frames

Verb		Class			K
produce	√	CREATE 92	PERFORM 6	OTHER 2	.73
take	√	PERFORM 10	STEAL 55	OTHER 33	.71
kick	√	HIT 33	BODY-INT. MOTION 2	OTHER 11	.78

Table 4.15: Semantic preferences for verbs with the transitive frame

Verb		Class				K
assess	√	ASSESS 95	PRICE 4	OTHER 1		.79
bang	√	HIT 35	S. EMISSION 58	OTHER 0		.93
bite	√	HURT 83	SWAT 12	OTHER 5		.89
clip	√	BRAID 23	CUT 22	OTHER 55		.67
crack	√	BRAID 54	S. EMISSION 11	OTHER 35		.78
crash	√	BREAK 57	S. EMISSION 8	OTHER 35		.78
demolish	√	AMUSE 38	DESTROY 57	OTHER 5		.81
draw		PULL 67	SCRIBBLE 26	OTHER 6		.77
feel	√	HUNT 0	SEE 56	OTHER 44		.67
file	√	BRAID 6	CARVE 3	POCKET 56	OTHER 35	.98
grind	√	BUILD 22	CARVE 3	CRANE 22	OTHER 53	.91
insult	√	AMUSE 49	JUDGMENT 31	OTHER 1		.90
kick	√	HIT 59	THROW 21	OTHER 19		.83
lick	√	TOUCH 61	W. MANNER 33	OTHER 6		.87
miss	√	ADMIRE 39	CONT. LOC. 59	OTHER 2		.75
move		ADMIRE 4	ROLL 1	SLIDE 75	OTHER 20	.73
poke	√	POKE 89	RUMMAGE 1	OTHER 10		.83

Table 4.16: Semantic preferences for verbs with the transitive frame

Verb		Class				K
push		CARRY 2	PUSH-PULL 76	OTHER 22		.83
rub	√	CRANE 72	W. MANNER 23	OTHER 4		.90
salute		CURTSEY 62	JUDGMENT 34	OTHER 4		.75
savour		ADMIRE 87	SIGHT 10	OTHER 2		.71
scrape	√	CUT 15	W. MANNER 49	OTHER 36		.90
scratch	√	HURT 67	W. MANNER 24	OTHER 9		.88
scrutinize	√	ASSESS 50	SIGHT 46	OTHER 3		.70
shoot	√	POISON 57	SWAT 5	THROW 12	OTHER 26	.81
stab	√	HIT 38	POISON 59	OTHER 3		.85
smash	√	BREAK 60	HIT 17	OTHER 22		74
study	√	ASSESS 46	LEARN 32	SIGHT 22	OTHER 0	.72
suck		CHEW 41	W. MANNER 49	OTHER 10		.69
support		ADMIRE 91	CONT. LOC. 9	OTHER 0		.81
toast	√	COOK 28	JUDGMENT 58	OTHER 2		.88
tug		CARRY 2	PUSH-PULL 83	OTHER 14		.81
value		ADMIRE 87	PRICE 13	OTHER 0		.75
whisk	√	SPANK 12	W. MANNER 61	OTHER 4		.83

TIGUOUS LOCATION, and PRICE which license only two frames. As a result, the parameter $P(f|c)$ masks the contribution of the parameter $P(c)$ and results in a counterintuitive ranking.

In general, the agreement on the class annotation task was good with Kappa values ranging from .66 to 1.00 (see Tables 4.12–4.16). The class OTHER in most cases was less frequent than the dominant verb class. For a few verbs a large number of tokens with erroneous frames were identified (see *take* in Table 4.12). In other cases, none of the senses listed in Levin (1993) corresponded to the dominant class of the verb. For example, most tokens for the verb *clip* mean "attach" or "pass rapidly" instead of CUT or BRAID (see Table 4.15). The annotation revealed that for most verbs the class distinctions adopted by Levin were attested in the corpus

data providing further empirical support for the proposed classification.

4.4.3. Discussion

In Experiment 6 we explored the degree to which corpus distributions in the form of syntactic frame and verb class information can be used to infer the dominant class of a polysemous verb in cases where syntactic information alone does not provide clues for disambiguation. In doing so, we cast the task of verb class disambiguation in a probabilistic framework which exploits Levin's semantic classification and frame frequencies acquired from the BNC. The approach is promising in that it achieves high precision with a simple model which has a straightforward interpretation in a Bayesian framework.

The semantic preferences which we generate can be thought of as default semantic knowledge, to be used in the absence of any explicit contextual or lexical semantic information to the contrary (see Tables 4.12–4.16). Consider the verb *write* for example. The model comes up with an intuitively reasonable ranking (see Table 4.12): we more often write things to people (MESSAGE TRANSFER reading) than for them (PERFORMANCE reading). However, faced with a sentence like *Max wrote Elisabeth a book*, pragmatic knowledge forces us to prefer the PERFORMANCE reading versus the MESSAGE TRANSFER reading. This contrasts with work in word sense disambiguation where each verb is disambiguated within its local context of occurrence. The semantic preferences we generate can serve as input to a word sense disambiguation procedure which defaults to the most probable meaning unless context indicates otherwise. Note that the model can be easily extended to incorporate other sources of information such as local context in order to inform the class selection process. Selectional restrictions is an obvious extension. Furthermore, beyond selectional restrictions, surface grammatical information pertaining to the verbal arguments can provide important clues for disambiguation. For example, a verb often taking a reflexive pronoun as its object is more likely to be a REFLEXIVE VERB OF APPEARANCE than a verb which never subcategorizes for a reflexive object.

Recall from the previous section that our model does not take the identity of the polysemous verb into account in order to derive a ranking for its meanings. Despite this limitation, our model can be used to automatically annotate a corpus with Levin compatible semantic classes which after manual inspection could provide better estimates for the initial model parameters. Adopting a bootstrapping approach (Hearst 1991; Yarowsky 1995) would allow us to derive increasingly accurate models of verb class ambiguity, while producing increasingly larger amounts of training data.

Finally, note that Experiment 6 focused solely on frames relating to the dative, benefactive, conative, and possessor object alternation. The number of frames we considered was too limited to address issues such as: (a) the relations between frames and classes (what are the frames for which the semantic class is predicted most accurately) and (b) the relations between verbs and classes (what are the verbs for which the semantic class is predicted most accurately).

Experimentation with a full scale subcategorization dictionary acquired from the BNC could potentially address these questions.

4.5. General Discussion

In this chapter we have presented a probabilistic model of verb class ambiguity based on Levin's (1993) semantic classification. Our results show that subcategorization information acquired automatically from corpora provides important cues for verb class disambiguation (see Experiment 5). In the absence of subcategorization cues, corpus-based distributions and quantitative approximations of linguistic concepts can be used to derive a preference ordering on the set of verbal meanings (see Experiment 6).

Our model relies on Levin's (1993) linguistic generalizations. We augment Levin's classification with probabilistic information which reduces verb class ambiguity by ranking potential meanings in terms of likelihood. Our results support the surface cueing approach put forward in Chapter 3. The model's parameters were estimated using simple distributions that can be easily extracted from corpora. Our heuristic approach yielded good results even though some of our assumptions were linguistically unmotivated.

Although our original aim was to develop a probabilistic framework which exploits Levin's (1993) linguistic classification and the systematic correspondence between syntax and semantics, a limitation of the model is that it cannot infer class information for verbs not listed in Levin. For these verbs $P(c)$, and hence $P(c,f,v)$, will be zero. Recent work in computational linguistics (e.g., Schütze 1993) and cognitive psychology (e.g., Landauer and Dumais 1997) has shown that large corpora implicitly contain semantic information, which can be extracted and manipulated in the form of co-occurrence vectors. One possible approach would be to compute the centroid (geometric mean) of the vectors of all members of a semantic class. Given an unknown verb (i.e., a verb not listed in Levin) we can decide its semantic class by comparing its semantic vector to the centroids of all semantic classes. For example, we could determine class membership on the basis of the closest distance to the centroid representing a semantic class (see Patel, Bullinaria, and Levy 1998 for a proposal similar in spirit). Another approach put forward by Dorr and Jones (1996) utilizes WordNet (Miller and Charles 1991) to find similarities (via synonymy) between unknown verbs and verbs listed in Levin. Once we have chosen a class for an unknown verb, we are entitled to assume that it will share the broad syntactic and semantic properties of that class.

In this chapter we demonstrated that the ambiguity exhibited by Levin's (1993) classification of verb meanings can be constrained via a model which combines Levin's inventory of meanings and the acquired corpus frequencies. In line with Chapter 3 we have shown that a shallow approach which uses linguistic theory on a par with corpus frequencies is useful for quantifying linguistic generalizations. Furthermore, the approach put forward in this chapter

can be also useful for practical applications that could benefit from a probabilistic ranking of the set of possible interpretations (for example by selecting the most likely one).

4.6. Related Work

Levin's (1993) taxonomy has been used in a variety of studies that aim to acquire lexical semantic information on the basis of the assumption that verbal meaning can be gleaned from corpora using cues pertaining to syntactic structure (Merlo and Stevenson 1999; Schulte im Walde 1998, 2000; Stevenson and Merlo 2000). Other work has used Levin's list of verbs (in conjunction with related lexical resources) for the creation of dictionaries that exploit the systematic correspondence between syntax and meaning (Dang et al. 1997; Dorr 1997; Dorr and Jones 1996). In this section we discuss lexicographic work based on Levin's taxonomy together with work related to the probabilistic approach advocated in this chapter.

Dang et al. (1997) aim to extend Levin's (1993) taxonomy by discovering relations between WordNet and Levin's semantic classification. In order to do this they augment Levin's existing semantic classes with a set of "intersective" classes which are created by grouping together sets of existing classes which share a minimum of three members. By creating intersective classes Dang et al. (1997) simulate a WordNet like semantic hierarchy by defining inter-class relations.

Most statistical approaches, including ours, treat verbal meaning assignment as a semantic classification task. The underlying question is the following: how can corpus information be exploited in deriving the semantic class for a given verb? Despite the unifying theme of using corpora and corpus distributions for the acquisition task, the approaches differ in the inventory of classes they employ, in the methodology used for inferring semantic classes and the specific assumptions concerning the verbs to be classified (i.e., can they be polysemous or not).

Merlo and Stevenson (1999) and Stevenson and Merlo (2000) use grammatical features (acquired from corpora) to classify verbs into three semantic classes: unergative, unaccusative, and object-drop. These classes are abstractions of Levin's (1993) classes and as a result yield a coarser classification. For example, object-drop verbs comprise a variety of Levin classes such as GESTURE verbs, CARING verbs, LOAD verbs, PUSH-PULL verbs, MEET verbs, SOCIAL INTERACTION verbs, AMUSE verbs, etc. Unergative, unaccusative, and object-drop verbs have identical subcategorization patterns (i.e., they alternate between the transitive and intransitive frame), yet distinct argument structures. This means that these three classes of verbs differ in the thematic roles they assign to their arguments. For example, when attested in the intransitive frame the subject of an object-drop verb is an Agent, whereas the subject of an unaccusative verb is a Theme. Under the assumption that differences in thematic role assignment uniquely identify semantic classes, numeric approximations of argument structure are derived from cor-

pora and used in a machine learning paradigm to classify verbs in their semantic classes.

More specifically, the features of transitivity, causativity, animacy of the subject, passive voice, and past participle are used as indicators of argument structure. For example, verbs with animate subjects are more likely to assign the thematic role of Agent rather than Theme. The approach is evaluated on 59 verbs manually selected from Levin (1993) (20 unergatives, 20 object-drop, and 19 unaccusatives). It is assumed that these verbs are monosemous, i.e., they can be either ergative, unergative or object-drop. The ACL/DCI corpus is used to obtain counts for transitivity, passive voice, and past participle. Causativity and animacy are approximated from a parsed version of the Wall Street Journal. A decision-tree learner trained on these features achieves a precision of 69.8% on the classification task over a chance baseline of 34%.

Schulte im Walde (1998, 2000) uses subcategorization information and selectional restrictions to cluster verbs into Levin compatible semantic classes. Subcategorization frames are induced from the BNC using a robust statistical parser (see Section 3.7 for details). The selectional restrictions are acquired using Resnik's (1993) information-theoretic measure of selectional association which combines distributional and taxonomic information (e.g., WordNet) in order to formalize how well a predicate associates with a given argument. Two sets of experiments are run to evaluate the contribution of selectional restrictions using two types of clustering algorithms, iterative clustering and latent class clustering (see Schulte im Walde 1998 for details). The approach is evaluated on 153 verbs taken from Levin, 53 of which are polysemous (i.e., belong to more than one class). The size of the derived clusters is restricted to four verbs and compared to Levin: verbs are classified correctly if they are members of a non-singleton cluster which is a subset of a Levin class. Polysemous verbs can be assigned to distinct clusters only using the latent class clustering method. The best results achieve a recall of 36% and a precision of 61% (over a baseline of 5%, calculated as the number of randomly created clusters which are subsets of a Levin class) using subcategorization information only and iterative clustering. Inclusion of information about selectional restrictions yields a lower accuracy of 38% (with a recall of 20%), again using iterative clustering.

Dorr and Jones (1996) use Levin's (1993) taxonomy to show that there is a predictable relationship between verbal meaning and syntactic behavior. Similarly to Schulte im Walde (1998, 2000) Dorr and Jones use information about the syntactic structure of a given verb with the aim to discover generalizations pertaining to meaning, although their approach is not statistical and does not make use of a large corpus. Dorr and Jones create a database of Levin's verb classes and the sentences exemplifying them (including both positive and negative examples, i.e., examples marked with asterisks). A parser is used to extract basic syntactic patterns for each semantic class. These patterns form the syntactic signature of the class. 97.9% of the semantic classes are identified uniquely by their syntactic signatures. Grouping verbs (instead of classes) with identical signatures to form a semantic class yields a 6.3% overlap with Levin classes. Their results are somewhat difficult to interpret since in practice information

about a verb and its syntactic signature is not available. Note that under Dorr and Jones's approach information about the syntactic signature of a verb is needed in order to classify it into a Levin class. Schulte im Walde's study shows that acquisition of syntactic signatures (i.e., subcategorization frames) from corpora is feasible, however these acquired signatures are not necessarily compatible with Levin and in most cases depart from those derived by Dorr and Jones.

Our work focuses on the ambiguity inherently present in Levin's (1993) classification. The problem is ignored by Stevenson and Merlo (2000) who focus only on monosemous verbs. Polysemous verbs are included in Schulte im Walde's (2000) experiments: the clustering approach can go so far as to identify more than one class for a given verb without, however, providing information about its dominant class. We recast Levin's taxonomy in a statistical framework and show in agreement with Stevenson and Merlo and Schulte im Walde that corpus-based distributions provide important information for semantic classification, especially in the case of polysemous verbs whose meaning cannot be easily inferred from the immediate surrounding context. Like Schulte im Walde, our approach uses subcategorization frames extracted from the BNC (although using a different methodology, see Chapter 3). We employ Levin's inventory of semantic classes arriving at a finer grained classification than Stevenson and Merlo.

Unlike Schulte im Walde (2000) and Stevenson and Merlo (2000), we ignore information about the arguments of a given verb either in the form of selectional restrictions or argument structure. This is clearly a limitation. However, such information can be easily incorporated in the model presented in Section 4.2 in the form of conditional probabilities where the verb is, for example, conditioned on the thematic role of its arguments (see also the discussion in Section 4.4.3). Unlike Stevenson and Merlo, Schulte im Walde and Dorr and Jones (1996) we provide a general probabilistic model which assigns a probability to each class of a given verb by calculating the probability of a complex expression in terms of the probability of simpler expressions that compose it. Our model can be easily extended to incorporate additional information in order to find the most expected meaning out of those allowed by available lexico-grammatical knowledge. In the context of this general probabilistic framework, our work can benefit by taking into account Stevenson and Merlo's thematic role distinctions as well as Schulte im Walde's full-scale subcategorization dictionary.

4.7. Summary

In this chapter we demonstrated that the ambiguity exhibited by Levin's (1993) classification of verb meanings can be constrained via a probabilistic model that combines Levin's inventory of meanings and corpus frequencies acquired from the BNC using the surface cueing approach. Our results show that frequency distributions of subcategorization frames within and across

classes can satisfactorily derive the most salient meaning for a polysemous verb in the absence of any explicit contextual or lexical semantic information to the contrary.

In the following chapter we turn to adjective-noun combinations and show how the proposed model (see Section 4.2) can be used to infer the meaning of polysemous adjectives. In contrast to the experiments reported here, where the verb meanings are provided by Levin, interpretations for polysemous adjectives are acquired automatically from the corpus. As in the case of polysemous verbs, the probabilistic model is used to provide a ranking of meanings, again exploiting systematic regularities between syntax and semantics.

Chapter 5

A Probabilistic Model of Adjective-Noun Ambiguity

In this chapter we further evaluate the probabilistic model introduced in the previous chapter by looking at polysemous adjective-noun combinations. More specifically, we concentrate on polysemous adjectives whose meaning varies depending on the noun they modify (e.g., *difficult*). In contrast to the previous chapter, where we used a linguistic classification (i.e., Levin 1993) as an inventory of the meanings of verbs in relation to their arguments, we derive the meanings of adjective-noun combinations directly from the corpus. The acquired meanings and their ranking are evaluated against human intuitions. We conduct an experiment which provides evidence that our model produces a preference ordering on the meanings of adjective-noun combinations which correlates reliably with human judgments.

5.1. Introduction

The semantic properties of adjectives have been extensively studied in the theoretical linguistics literature. Adjectives exhibit a widely polymorphic behavior, aspects of which several semantic classifications have attempted to capture. A well-known classification of adjectives is based on their logical behavior and divides adjectives into three classes: *extensional, intensional*, and *scalar* (Chierchia and McConnell-Ginet 1990). Extensional adjectives denote properties; when they modify a noun the meaning of the adjective-noun phrase is the intersection of the semantics contributed by the noun and the adjective. *Red* is a typical example of an extensional adjective: a *red dress* is interpreted as the intersection of the set of red things and the set of dresses. Intensional adjectives are property-modifying (i.e., they denote a function from properties to properties). For example, the adjective-noun phrase *former president* does not denote the individual that is a president and former; instead, it denotes the individual that was president in a preceding term. Scalar adjectives denote properties relative to a norm or a

standard of comparison. For example, a *small elephant* is small for an elephant or small as elephants go (i.e., it is not a small animal).

Adjectives are also classified in terms of their gradability (Lyons 1977; Quirk et al. 1985). *Gradable* adjectives denote a property that can vary by degrees (e.g., *deep, fast, big*). Gradability can be indicated by the use of degree modifiers (e.g., *very, much, highly, extremely*) and by comparison markers (i.e., the addition of the suffixes *-er* and *-est* or by modification with *more* and *most*). The modifier locates the adjective on a scale of comparison, at a position higher or lower than the one indicated by the adjective alone. Adjectives denoting the highest position on a scale are non-gradable (e.g., *dead, principal, pregnant*). Note that an adjective can be extensional and gradable (e.g., an *extremely red dress*) or scalar and gradable (e.g., a *very small elephant*).

An important semantic relation holding between pairs of adjectives is antonymy, i.e., semantic opposition. Antonymy is the basis for the semantic organization of adjectives in WordNet (Miller et al. 1990). Related to antonymy are the semantic *orientation* and *markedness* of adjectives (Lyons 1977). The orientation usually indicates whether the adjective receives a positive or negative interpretation. For example, the adjectives *intelligent* and *simple* have a positive orientation, whereas the adjectives *stupid* and *simplistic* have a negative orientation. Given two contrasting adjectives the *unmarked* adjective denotes a generic property without explicitly making reference to a norm or a standard (e.g., the adjective *tall* in the question *How tall is Peter?*); the *marked* adjective denotes a property that deviates from the norm (e.g., the adjective *short* in the question *How short is Peter?*). Note that the orientation of an adjective is highly dependent on contextual or pragmatic factors. For example, *simple* can have a positive orientation when contrasted to *complicated* and a negative orientation when contrasted to *elegant.*

Semantically, adjectives, more than other categories, are able to take on different meanings depending on the noun they modify (Lahav 1989; Pustejovsky 1995; Sapir 1944; Vendler 1968). Consider the adjective *red* for example. A *red ball* is ball which is colored red, a *red party* is a left-wing party, and a *red pen* is pen which writes red. Adjectives which denote physical properties behave similarly: a *hot car* has a hot engine (or perhaps a hot interior), a *hot dish* is spicy, and *hot water* is warm. Adjectives like *difficult, easy, fast,* or *good* pattern with *hot* and *red* in that they receive different meanings when modifying different nouns. Furthermore, they display different meanings even when they modify a single noun: these adjectives are ambiguous across and within the nouns they modify. Consider the examples in (5.1). The meaning of the adjective *fast* varies depending on whether it modifies the nouns *programmer, scientist,* or *plane.* A *fast programmer* is typically "a programmer who programs quickly", a *fast plane* is typically "a plane that flies quickly", a *fast scientist* can be "a scientist who publishes papers quickly", "who performs experiments quickly", "who observes something quickly", "who reasons, thinks, or runs quickly". Similarly, a *fast plane* is not only "a plane that flies quickly",

but also "a plane that lands, takes off, turns, or travels quickly". Even the more restrictive *fast programmer* allows more than one interpretation. As shown in (5.3), taken from Lascarides and Copestake (1998: 394), the discourse context triggers the interpretation of "a programmer who runs fast".

(5.1) a. fast programmer
b. fast plane
c. fast scientist

(5.2) a. easy problem
b. difficult language
c. good cook
d. good soup

(5.3) a. All the office personnel took part in the company sports day last week.
b. One of the programmers was a good athlete, but the other was struggling to finish the courses.
c. The fast programmer came first in the 100m.

Adjectives like *fast* have been extensively studied in the lexical semantics literature (Bouillon 1997; Lahav 1989; Pustejovsky 1995) and their properties have been known at least since Vendler (1968). The meaning of adjective-noun combinations like those in (5.1) and (5.2) are usually paraphrased with a verb modified by the adjective in question or its corresponding adverb. For example, an *easy problem* is "a problem that is easy to solve" or "a problem that one can solve easily". As Vendler (1968: 92) points out in most cases not one verb, but a family of verbs is needed to account for the meaning of adjective-noun combinations like those in (5.1) and (5.2). Vendler further observes that the noun figuring in an adjective-noun combination is usually the subject or object of the paraphrasing verb. Although the adjective *fast* usually triggers a verb-subject interpretation (see the examples in (5.1)), the adjectives *easy* and *difficult* trigger a verb-object interpretation (see the examples in (5.2a,b)). An *easy problem* is usually "a problem that is easy to solve", whereas a *difficult language* is "a language that is difficult to learn, speak, write, or understand". Adjectives like *good* allow either verb-subject or verb-object interpretations: a *good cook* is "a cook who cooks well", whereas *good soup* is "soup that tastes good" or "soup that is good to eat".

The polysemy of adjectives like *fast* or *easy* has led Pustejovsky (1995) to argue against lexicons which describe lexical meaning simply by sense enumeration. Deriving the meaning of adjective-noun constructions like (5.1) and (5.2) within a sense enumerative framework means that distinct senses have to be provided for each noun or, more generally, for each noun class the adjective modifies. Pustejovsky's account of adjectival polysemy relies on the fact that the meaning of adjectives like *easy* is determined largely by the semantics of the noun they modify. Pustejovsky assumes that nouns have a *qualia structure* as part of their lexical

entries, which among other things, specifies possible events associated with the entity. For example, the telic (purpose) role of the qualia structure for *problem* has a value equivalent to *solve*. Adjectives can be seen as modifying only one or a subset of the qualia for a noun. The adjective *easy* is an event predicate, i.e., it selectively modifies the events associated with the nouns it is in construction with. When *easy* is combined with *problem*, it predicates over the telic role of *problem* and consequently the adjective-noun combination receives the interpretation "a problem that is easy to solve". Pustejovsky calls the semantic process of selecting and operating on a specific substructure of a lexical entry *Selective Binding*. Note that in Pustejovsky's framework the polysemy of adjectives like *easy* is accounted for by lexical processes operating on lexical entries, thus avoiding the proliferation of senses via enumeration.

Pustejovsky (1995) does not give an exhaustive list of the telic roles a given noun may have. In contrast to Vendler (1968), who acknowledges the fact that adjective-noun combinations like the ones in (5.1) and (5.2) trigger more than one interpretation (in other words, there may be more than one possible event associated with the noun modified by the adjective in question), Pustejovsky implicitly assumes that nouns or noun classes have one—perhaps default—telic role. Although the number of possible interpretations for adjective-noun combinations like *fast scientist* are virtually unlimited, some interpretations are more likely than others. Out of context, *fast scientist* is more likely to be interpreted as "a scientist who performs experiments quickly" or "who publishes quickly" rather than as "a scientist who draws or drinks quickly".

In this chapter we focus on polysemous adjective-noun combinations (see (5.1) and (5.2)) and attempt to address the following questions: (a) Can the meanings of these adjective-noun combinations be acquired automatically from corpora? (b) Can we constrain the number of interpretations by providing a ranking on the set of possible meanings? (c) Can we determine if an adjective has a preference for a verb-subject or verb-object interpretation? We provide a probabilistic model (based on the model introduced in Chapter 4) which combines distributional information about how likely it is for any verb to be modified by the adjective in the adjective-noun combination or its corresponding adverb with information about how likely it is for any verb to take the modified noun as its object or subject. As in Chapter 4, we obtain quantitative information about verb-adjective modification and verb-argument relations from the BNC via partial parsing. Our results not only show that we can predict meaning differences when the same adjective modifies different nouns, but we can also derive—taking into account Vendler's (1968) observation—a cluster of meanings for a single adjective-noun combination.

We evaluate our results by comparing the model's predictions against human judgments and show that the model's ranking of meanings correlates reliably with human intuitions: meanings that are found highly probable by the model are also rated as plausible by the subjects. Furthermore, we demonstrate that the model's predictions can be used to arrive at a tripartite distinction of adjectives depending on the type of paraphrase they prefer: subject-biased

adjectives tend to modify nouns which act as subjects of the paraphrasing verb, object-biased adjectives tend to modify nouns which act as objects of the paraphrasing verb, whereas equi-biased adjectives display no preference for either argument role. We show that the argument preferences predicted by the model correspond to preferences displayed by humans.

In the following sections we present our probabilistic model of adjective-noun ambiguity and describe the model parameters (see Section 5.2). In Experiment 7 we demonstrate the properties of the model using examples from the literature (see Section 5.3). In Experiment 8 we use the model to derive the meaning paraphrases for adjective-noun combinations randomly selected from the BNC (see Section 5.4) and formally evaluate the results against human intuitions (see Section 5.4.2). Finally, in Experiment 9 we demonstrate that when compared against human judgments our model outperforms a naive baseline in deriving a preference ordering for the meanings of polysemous adjective-noun combinations (see Section 5.5).

5.2. The Model

5.2.1. Formalization of Adjective-Noun Polysemy

Consider again the adjective-noun combinations in (5.1) and (5.2). In order to come up with the meaning of "plane that flies quickly" for *fast plane* we would like to find in the corpus a sentence whose subject is the noun *plane* or *planes* and whose main verb is *fly*, which in turn is modified by the adverb *fast* or *quickly*. In the general case we would like to find in the corpus sentences indicating what planes do fast. Similarly, for the adjective-noun combination *fast scientist* we would like to find in the corpus information indicating what the activities that scientists perform fast are, whereas for *easy problem* we need information about what one can do with problems easily (e.g., one can solve problems easily) or about what problems are (e.g., easy to solve or set).

In sum, in order to come up with a paraphrase of the meaning of an adjective-noun combination we need to know which verbs take the head noun as their subject or object and are modified by an adverb corresponding to the modifying adjective. This can be expressed as the joint probability $P(a,n,rel,v)$ where v is the verbal predicate modified by the adverb a (directly derived from the adjective present in the adjective-noun combination) bearing the argument relation *rel* (i.e., subject or object) to the head noun n. By choosing the ordering $\langle v,n,a,rel\rangle$ for the variables a,n,rel, and v we can rewrite $P(a,n,rel,v)$ (using the chain rule) as follows:

$$P(a,n,rel,v) = P(v) \cdot P(n|v) \cdot P(a|v,n) \cdot P(rel|v,n,a) \tag{5.4}$$

Although the parameters $P(v)$ and $P(n|v)$ can be straightforwardly estimated from the BNC, the estimation of $P(a|v,n)$ and $P(rel|v,n,a)$ is somewhat problematic. Let us consider more

closely the term $P(rel|v,n,a)$ which can be estimated as shown in (5.5) below.

$$(5.5) \quad P(rel|v,n,a) = \frac{f(v,n,a,rel)}{f(v,n,a)}$$

One way to estimate $f(v,n,a,rel)$ would be to fully parse the corpus so as to identify the verbs which take the head noun n as their subject or object and are modified by the adverb a. For the adjective-noun combination *fast plane* there are only six sentences in the entire BNC that could be used for the estimation of $f(v,n,a,rel)$ (see the examples in (5.6)). According to these sentences the most likely interpretation for *fast plane* is "a plane that goes fast" (see examples (5.6a)–(5.6c)). The interpretations "plane that swoops in fast", "plane that drops down fast" and "plane that flies fast" are all equally likely, since they are attested in the corpus only once (see examples (5.6d)–(5.6f)). This is rather unintuitive since *fast planes* are more likely to fly than swoop in fast. For the adjective-noun combination *fast programmer* there is only one sentence relevant for the estimation of $f(v,n,a,rel)$ in which the modifying adverbial is not *fast* but the semantically related *quickly* (see example (5.7)). The sparse data problem carries over to the estimation of the frequency $f(v,n,a)$.

(5.6)
a. The plane went so fast it left its sound behind.
b. And the plane's going slightly faster than the Hercules or Andover.
c. He is driven by his ambition to build a plane that goes faster than the speed of sound.
d. Three planes swooped in, fast and low.
e. The plane was dropping down fast towards Bangkok.
f. The unarmed plane flew very fast and very high.

(5.7) It means that programmers will be able to develop new applications more quickly.

We avoid these estimation problems by reducing the parameter space. In particular, we make the following independence assumptions:

$$(5.8) \quad P(a|v,n) \approx P(a|v)$$

$$(5.9) \quad P(rel|v,n,a) \approx P(rel|v,n)$$

We assume that the likelihood of seeing an adverbial modifying a verb bearing an argument relation to a noun is independent of that specific noun (see equation (5.8)). In other words we only estimate the likelihood of a verb to be modified by a particular adverb or adjective. Accordingly, we assume that the likelihood of the argument relation *rel* given a verb v that takes a noun n as its subject or object and is modified by an adverb a is independent of the

adverb a. By substituting (5.8) and (5.9) into (5.4), $P(a,n,rel,v)$ can be written as:

$$P(a,n,rel,v) \approx P(v) \cdot P(n|v) \cdot P(a|v) \cdot P(rel|v,n) \tag{5.10}$$

We estimate the probabilities $P(v)$, $P(n|v)$, $P(a|v)$, and $P(rel|v,n)$ as follows:

$$P(v) = \frac{f(v)}{\sum_i f(v_i)} \tag{5.11}$$

$$P(n|v) = \frac{f(n,v)}{f(v)} \tag{5.12}$$

$$P(a|v) = \frac{f(a,v)}{f(v)} \tag{5.13}$$

$$P(rel|v,n) = \frac{f(rel,v,n)}{f(v,n)} \tag{5.14}$$

By substituting equations (5.11)–(5.14) into (5.10) and simplifying the relevant terms, (5.10) is rewritten as follows:

$$P(a,n,rel,v) \approx \frac{f(rel,v,n) \cdot f(a,v)}{f(v) \cdot \sum_i f(v_i)} \tag{5.15}$$

Assume we want to discover a meaning paraphrase for the adjective-noun combination *fast plane*. We need to find the verb or verbs v and the relation *rel* (i.e., subject or object) that maximize the term $P(fast,plane,rel,v)$. Table 5.1 gives a list of the most frequent verbs modified by the adverb *fast* in the BNC (see the term $f(a,v)$ in equation (5.15)), whereas Table 5.2 lists the verbs for which the noun *plane* is the most likely object or subject (see the term $f(rel,v,n)$ in equation (5.15)). We describe how the frequencies $f(rel,v,n)$, $f(a,v)$, and $f(v)$ were estimated from a lemmatized version of the BNC in the following section.

Table 5.1 can be thought of as a list of the activities that can be fast (i.e., going, growing, flying), whereas Table 5.2 specifies the events associated with the noun *plane*. Note that despite our simplifying assumptions the model given in (5.14) will come up with plausible meanings for adjective-noun combinations like *fast plane*. Note that the verbs *fly*, *come*, and *go* are most likely to take the noun *plane* as their subject (see Table 5.2). These verbs also denote activities that are fast (see Table 5.1). Further note that a subject interpretation is more likely than an object interpretation for *fast plane* since none of the verbs likely to have *plane* as their object are modified by the adverb *fast* (compare Tables 5.1 and 5.2).

Table 5.1: Most frequent verbs modified by the adverb *fast*

v	$f(fast,v)$	v	$f(fast,v)$
go	29	work	6
grow	28	grow in	6
beat	27	learn	5
run	16	happen	5
rise	14	walk	4
travel	13	think	4
move	12	keep up	4
come	11	fly	4
drive	8	fall	4
get	7	disappear	4

Table 5.2: Most frequent verbs taking as an argument the noun *plane*

v	$f(\text{SUBJ},v,plane)$	v	$f(\text{OBJ},v,plane)$
fly	20	catch	24
come	17	board	15
go	15	take	14
take	14	fly	13
land	9	get	12
touch	8	have	11
make	6	buy	10
arrive	6	use	8
leave	5	shoot	8
begin	5	see	7

5.2.2. Parameter Estimation

We estimated the parameters of the model outlined in the previous section from a part-of-speech tagged and lemmatized version of the BNC. The estimation of the terms $f(v)$ and $\sum_i f(v_i)$ (see (5.15)) reduces to the number of times a given verb is attested in the corpus. In order to estimate the terms $f(rel,v,n)$ and $f(a,v)$ the corpus was automatically parsed by Cass (Abney 1996), a robust chunk parser designed for the shallow analysis of unrestricted text (for a detailed description of the parser see Section 2.3.2 in Chapter 2). Section 5.2.2.1 details how the frequencies $f(rel,v,n)$ and $f(v,n)$ were acquired.

5.2.2.1. The parser

For the estimation of $f(rel,v,n)$ we used the parser's built-in function to extract tuples of verb-subjects and verb-objects (see the examples in (5.16)). The tuples obtained from the parser's output are an imperfect source of information about argument relations. Bracketing errors as well as errors in identifying chunk categories accurately result in extracting tuples whose lex-

ical items do not stand in a verb-argument relationship. For example, the verb is missing from tuples (5.17a,b) (*people* and *whose* are identified as subjects of *isolated* and *behalf*, respectively), the noun is missing from tuples (5.17c,d) (the adverb *there* is the subject of *drink*, the adjective *good* is the subject of *smile*), and both the verb and the noun are missing from tuple (5.17e) (where the relative pronoun *who* is the subject of the adjective *ill*).

(5.16)	a.	change situation	SUBJ
	b.	analyse participant	SUBJ
	c.	come off heroin	OBJ
	d.	appear on screen	OBJ
	e.	deal with situation	OBJ
(5.17)	a.	isolated people	SUBJ
	b.	behalf whose	SUBJ
	c.	drink there	SUBJ
	d.	smile good	SUBJ
	e.	ill who	SUBJ
(5.18)	a.	alten aus	SUBJ
	b.	rolex symbol	SUBJ

In order to compile a comprehensive count of verb-argument relations we tried to eliminate from the parser's output tuples containing erroneous verbs and nouns like those in (5.17). We did this by matching the verbs contained in the tuples against a list of all words tagged as verbs, and the nouns in the tuples against a list of all nouns in the BNC. Tuples containing words not included in the list were discarded. Furthermore, tuples containing verbs or nouns attested in a verb-argument relationship only once were also discarded, since they were mostly tagging or parsing mistakes. See the examples in (5.18) where *alten* and *rolex* are tagged as verbs (instead of nouns) and *aus* is mistakenly tagged as a noun. Finally, non-auxiliary instances of the verb *be* (e.g., be embassy OBJ, be prawn SUBJ) were eliminated since they contribute no semantic information with respect to the events or states that are possibly associated with the noun with which the adjective is combined.

Particle verbs (see (5.16c)) were included in verb-subject and verb-object tuples only if the particle was adjacent to the verb. Verbs followed by the preposition *by* and a head noun were extracted and counted as instances of verb-subject relations. The verb-object tuples also included prepositional objects (see (5.16d,e)). It was assumed that PPs adjacent to the verb headed by either of the prepositions *in, to, for, with, on, at, from, of, into, through, upon* were prepositional objects. This resulted in 737,390 types of verb-subject pairs and 1,078,053 types of verb-object pairs (see Table 5.3 which contains information about the tuples extracted from the corpus before and after the filtering).

Generally speaking, the frequency $f(a,v)$ represents not only a verb modified by an adverb derived from the adjective in question (see example (5.19a)) but also constructions like

Table 5.3: Tuples extracted from the BNC

	Tokens		Types		
Relation	Parser	Filtering	Tuples	Verbs	Nouns
SUBJECT	4,759,950	4,587,762	737,390	14,178	25,900
OBJECT	3,723,998	3,660,897	1,078,053	12,026	35,867

the ones shown in (5.19b,c), where the adjective takes an infinitival VP complement whose logical subject can be realized as a *for*-PP (see example (5.19c)). In cases of verb-adverb modification we assume access to morphological information which specifies what counts as a valid adverb for a given adjective. In most cases adverbs are formed by adding the suffix *-ly* to the base of the adjective (e.g., *slow-ly*, *easy-ly*). Some adjectives have identical adverbs (e.g., *fast*, *right*). Others have idiosyncratic adverbs (e.g., the adverb of *good* is *well*). It is relatively straightforward to develop an automatic process which maps an adjective to its corresponding adverb, modulo exceptions and indiosyncracies, however in the experiments described in the following sections this mapping was manually specified.

(5.19) a. comfortable chair → a chair *on* which one *sits comfortably*
b. comfortable chair → a chair that is *comfortable* to *sit on*
c. comfortable chair → a chair that is *comfortable* for me to *sit on*

Note that in cases where the adverb does not immediately succeed the verb (see sentence (5.7)) the parser is not guaranteed to produce a correct analysis due to its simple strategy of leaving ambiguities unattached. In order to estimate the frequency $f(a,v)$ we only looked at instances where the verb and the adverbial phrase modifying it (AdvP) were adjacent. More specifically, in cases where the parser identified an AdvP following a VP, we extracted the verb and the head of the AdvP (see examples (5.20b), (5.21b), (5.22c)). In cases where the AdvP was not explicitly identified we extracted the verb and the adverb immediately following or preceding it (see examples (5.20a), (5.21a), (5.22a), and (5.22b)) assuming that the verb and the adverb stand in a modification relation. The examples below illustrate the parser's output and the information that was extracted for the estimation of the quantity $f(a,v)$.

(5.20) a. [NP Some art historians] [VP write] well [PP about the present.] write well
b. [NP Oriental art] [VP came] [AdvP more slowly.] come slowly
(5.21) a. [NP The accidents] [VP could have been easily avoided.] avoid easily
b. [NP The issues] [VP will not be resolved] [AdvP easily.] resolve easily
(5.22) a. [NP A system of molecules] [VP is easily shown] [VP to stay constant.] show easily
b. [NP Their economy] [VP was so well run.] run well
c. [NP Arsenal] [VP had been pushed] [AdvP too hard.] push hard

Adjectives with infinitival complements (see (5.19b,c)) were acquired as follows: we concentrated only on adjectives immediately followed by infinitival complements with an op-

tionally intervening *for*-PP (see (5.19c)). The adjective and the main verb of the infinitival complement were counted as instances of the quantity $f(a,v)$. The examples in (5.23) illustrate the process.

(5.23) a. [NP These early experiments] [VP were easy] [VP to interpret.] easy interpret
b. [NP It] [VP is easy] [PP for an artist] [VP to show work independently.] easy show
c. [NP It] [VP is easy] [VP to show] [VP how the components interact.] easy show

Finally, the frequency $f(a,v)$ collapsed the counts from cases where the adjective was followed by an infinitival complement (see the examples in (5.23)) and cases where the verb was modified by the adverb corresponding to the related adjective (see the examples in (5.20)–(5.22)). For example, assume that we are interested in the frequency $f(easy, show)$. In this case, we will take into account not only sentences (5.23b,c) but also sentence (5.22a). Assuming this was the only evidence in the corpus, the frequency $f(easy, show)$ would be three.

Once we have obtained the frequencies $f(a,v)$ and $f(rel,v,n)$ we can determine what the most likely interpretations for a given adjective-noun combination are. Depending on the data (noisy or not) and the task at hand we may choose to estimate the probability $P(a,n,rel,v)$ from reliable corpus frequencies only (e.g., $f(a,v) > 1$ and $f(rel,v,n) > 1$). If we know the interpretation preference of a given adjective (i.e., subject or object), we may vary only the term v in $P(a,n,rel,v)$, keeping the terms n, a and rel constant. Alternatively, we could acquire the interpretation preferences automatically by varying both the terms rel and v. In Experiment 8 (see Section 5.4) we acquire both meanings and argument preferences for polysemous adjective-noun combinations.

In what follows we explain the properties of the model by applying it to a small number of adjective-noun combinations taken from the lexical semantics literature (i.e., Pustejovsky 1995 and Vendler 1968). We show that our model predicts variation in meaning when the same adjective modifies different nouns, and furthermore that it provides a fairly intuitive ranking of meanings for a given adjective-noun combination. We apply the model to examples randomly selected from the BNC in Experiment 8 and evaluate its performance against human judgments (see Section 5.4).

5.3. Experiment 7: Comparison against the Literature

5.3.1. Method

We selected 15 adjective-noun combinations discussed in the lexical semantics literature (Pustejovsky 1995; Vendler 1968). The adjective-noun combinations and their respective interpretations are given in Table 5.4. Note that although in some cases more than one meaning is provided (e.g., a *difficult language* is "a language that is difficult to speak, learn, write, or

Table 5.4: Paraphrases for adjective-noun combinations taken from the literature

good knife	→ a knife that cuts well	(Pustejovsky 1995:43)
good meal	→ a tasty meal (a meal that tastes good)	(Pustejovsky 1995:43)
good umbrella	→ an umbrella that functions well	(Pustejovsky 1995:43)
good poet	→ a poet who writes poems well	(Vendler 1968:101)
good shoe	→ a shoe that is good for wearing, for walking	(Vendler 1968:99)
fast boat	→ a boat driven quickly/a boat that is inherently fast	(Pustejovsky 1995:44)
fast game	→ the motions involved in the game are rapid and swift	(Pustejovsky 1995:44)
fast decision	→ a decision which takes a short amount of time	(Pustejovsky 1995:44)
fast horse	→ a horse that runs fast	(Vendler 1968:92)
easy problem	→ a problem that is easy to solve	(Vendler 1968:97)
easy planet	→ a planet that is easy to observe	(Vendler 1968:99)
easy text	→ text that reads easily	(Vendler 1968:99)
difficult language	→ a language that is difficult to speak, learn, write, understand	(Vendler 1968:99)
careful scientist	→ a scientist who observes, performs, runs experiments carefully	(Vendler 1968:92)
comfortable chair	→ a chair on which one sits comfortably	(Vendler 1968:98)

understand") in most cases a single interpretation is given (e.g., a *good knife* is "a knife that cuts well", an *easy text* is a "text that reads easily", etc.). We derived paraphrases for each adjective-noun combination in Table 5.4 using the probabilistic model outlined in Section 5.2. The model's parameters were estimated as explained in Section 5.2.2. No thresholds were employed for the frequencies $f(a,v)$ and $f(rel,v,n)$. Recall that the frequency $f(a,v)$ collapses the counts of adjectives co-occurring with infinitival complements and verbs modified by adverbs. We compiled counts corresponding to verb-adverb modification by mapping the adjective *good* to the adverbs *good* and *well*, the adjective *fast* to the adverb *fast*, *easy* to *easily* and *comfortable* to *comfortably*. The adverbial function of the adjective *difficult* is expressed only periphrastically (i.e., in a difficult manner, with difficulty). As a result we obtained the frequency $f(difficult,v)$ only on the basis of infinitival constructions (see the examples in (5.23)). Table 5.5 gives the five most likely interpretations for each adjective-noun combination (where v_1 is the most likely interpretation, v_2 is the second most likely interpretation, etc.).

5.3.2. Results

Let us now consider in more detail the interpretations the model comes up with. Pustejovsky (1995) suggests the interpretation "a knife that cuts well" for the adjective-noun combination *good knife*. This is the second most likely interpretation according to our probabilistic model (see Table 5.5). The model acquires additional less plausible meanings such as "a knife that goes well", "a knife that comes well", "a knife that takes something well", "a knife that buys something well". Although Pustejovsky focuses on a subject-related interpretation for *good knife*, the model also derives object-related interpretations: a *good knife* is "a knife that is good to use, to hold, to know, to draw, to take". The interpretations are fairly plausible with the exception perhaps of the paraphrase "a knife that is good to draw".

Table 5.5: Model-derived paraphrases for adjective-noun combinations, ranked in order of likelihood

$P(a,n,rel,v)$	v_1	v_2	v_3	v_4	v_5
$P(good,knife,\text{SUBJ},v)$	go	cut	come	take	buy
$P(good,knife,\text{OBJ},v)$	use	hold	know	draw	take
$P(good,meal,\text{SUBJ},v)$	go	cook	serve	choose	miss
$P(good,umbrella,\text{SUBJ},v)$	cover				
$P(good,umbrella,\text{OBJ},v)$	keep	wave	hold	run for	leave
$P(good,poet,\text{SUBJ},v)$	write	know	see	say	express
$P(good,shoe,\text{OBJ},v)$	wear	keep	buy	get	stick
$P(fast,boat,\text{SUBJ},v)$	travel	sink	go	come	disappear
$P(fast,boat,\text{OBJ},v)$	travel in	sink	drive	catch	get
$P(fast,game,\text{SUBJ},v)$	go	run	come	spread	start
$P(fast,decision,\text{OBJ},v)$	make	grow in	drive	avoid	get
$P(fast,horse,\text{SUBJ},v)$	run	learn	go	come	rise
$P(easy,problem,\text{OBJ},v)$	solve	deal with	identify	tackle	handle
$P(easy,planet,\text{OBJ},v)$	predict	identify	plunder	see on	work with
$P(easy,text,\text{OBJ},v)$	read	handle	use	interpret	understand
$P(difficult,language,\text{OBJ},v)$	understand	interpret	learn	use	speak
$P(careful,scientist,\text{SUBJ},v)$	calculate	proceed	investigate	study	analyse
$P(comfortable,chair,\text{OBJ},v)$	sink into	sit on	lounge in	relax in	nestle in

Consider now the pair *good meal* whose intuitive interpretation is "a meal that tastes good" (see Table 5.4). Although the model does not derive this particular interpretation, it derives complementary meanings such as "a meal that goes well", "a meal that cooks well", "a meal that serves well" (see Table 5.5). Similarly, although the model does not discover the suggested interpretation for the pair *good umbrella* (i.e., "an umbrella that functions well") it comes up with a plausible meaning (i.e., "an umbrella that covers well"). In fact, the meaning the model suggests can be considered as a subtype of the meaning suggested by Pustejovsky (1995): an umbrella functions well if it opens well, closes well, covers well, etc. Note also that the model derives object-related interpretations for *good umbrella*: "an umbrella that is good to keep, good for waving, good to hold, good to run for, good to leave". The meaning paraphrases are fairly plausible with the exception perhaps of the latter one.

The model and Vendler (1968) agree in their interpretation of the pairs *good poet* and *good shoe*. A *good poet* is "a poet who writes well", whereas a *good shoe* is "shoe that is good to wear" (see Table 5.5). The model further acquires the fairly plausible meanings "a poet who expresses himself well" for *good poet* and " a shoe that is good to keep, to buy, and get" for *good shoe*. Our model also comes up with plausible interpretations for the combinations *fast boat*, *fast game*, *fast decision*, and *fast horse*. In fact, the interpretations ranked as most likely by the model are similar to the ones proposed in the lexical semantics literature. A *fast boat* is "a boat that travels fast" according to the model; Pustejovsky's (1995) interpretation is semantically close ("a boat that is inherently fast"). Also notice the object-related interpretations derived by the model for *fast boat* (see Table 5.5). A *fast game* is "a game that goes or runs fast"

according to our model; Pustejovsky's proposal is semantically related: "the motions involved in the game are rapid and swift". The model correctly interprets a *fast decision* as "a decision that is fast to make" (see Pustejovsky's interpretation: "a decision which takes a short amount of time") and a *fast horse* as "a horse that runs fast" (see Vendler's identical interpretation in Table 5.4). Note further that according to the model a *fast horse* is not only "a horse that runs fast" but also a "horse that learns, goes, comes, and rises fast".

Similarly, an *easy problem* is not only "a problem that is easy to solve" (see Vendler's 1968 identical interpretation in Table 5.4) but also "a problem that is easy to deal with, identify, tackle, and handle" (see Table 5.5). The meaning of *easy problem* is different from the meaning of *easy text* which in turn is "easy to read, handle, interpret, and understand". The interpretations the model arrives at for the adjective-noun combination *difficult language* are a superset of the interpretations suggested by Vendler (see Table 5.4). The model comes up with the additional meanings "language that is difficult to interpret" and "language that is difficult to use". Although the meanings acquired by the model for *careful scientist* do not overlap with the ones suggested by Vendler (see Table 5.4) they seem intuitively plausible: a *careful scientist* is a "scientist who calculates, proceeds, investigates, studies, and analyses carefully". These are all possible events associated with scientists. Finally, note that the meanings derived for *comfortable chair* are also fairly plausible (the second most likely meaning is the one suggested by Vendler, see Table 5.4).

5.3.3. Discussion

Experiment 7 presented an example of how the probabilistic model outlined in Section 5.2 can be used to discover the meanings of adjective-noun combinations taken from the lexical semantics literature. Our probabilistic model combines distributional information about how likely it is for a verb to be modified by the adjective or adverb derived from the adjective present in an adjective-noun combination with information about how likely it is for any verb to take the related noun as its object or subject. We obtained quantitative information about verb-adjective and verb-adverb modification as well as verb-argument relations from the BNC via partial parsing. The results of Experiment 7 indicate that the probabilistic model can be used not only to predict different meanings when the same adjective modifies different nouns but also to derive a cluster of meanings for a single adjective-noun combination.

Although the model can be used to provide several interpretations for a given adjective-noun combination, not all of these interpretations are useful or plausible. Experiment 7 showed that meanings with top-ranked probabilities are intuitively plausible, although in some cases implausible meanings are also assigned top-ranked probabilities (for example, "a knife that buys something well" is ranked as the fifth most likely meaning for *good knife*, see Table 5.5). Furthermore, we did not explore the status of meanings with low probabilities. For an ideal model one would expect that top-ranked probabilities correspond to plausible meanings and

bottom-ranked probabilities correspond to implausible meanings. We test this prediction in Experiment 8. Another objection to the examples given in Tables 5.4 and 5.5 is that they may not be entirely representative of the types of polysemous adjective-noun combinations occurring in unrestricted text since they are taken from linguistic texts where emphasis is given on explaining the phenomenon at hand and the selected examples are typically straightforward illustrations of polysemous adjective-noun combinations. In other words, the adjective-noun combinations discussed in the previous section may be too easy for the model to handle. In Experiment 8 (see Section 5.4) we test our model on polysemous adjective-noun combinations randomly sampled from the BNC and formally evaluate our results against human judgments.

Finally, note that the meanings acquired by our model are a simplified version of the ones provided in the lexical semantics literature. In particular, note that an adjective-noun combination may be paraphrased with another adjective-noun combination (see Table 5.4 where *good meal* is paraphrased as "a tasty meal") or with a an NP instead of an adverb (see the paraphrase of *fast decision* in Table 5.4). We are making the simplifying assumption that a polysemous adjective-noun combination can be paraphrased by a sentence consisting of a verb whose argument is the noun the adjective is in construction with.

The probabilistic model discussed in the previous sections acquires meanings for polysemous adjective-noun combinations out of context. The derived meanings can be thought of as default semantic information associated with a particular adjective-noun combination. This means that our model is unable to predict the meaning of *fast programmer* when embedded in a context like the one given in (5.3).

5.4. Experiment 8: Comparison against Human Judgments

5.4.1. Method

The ideal test of the proposed model of adjective-noun polysemy will be with randomly chosen materials. We evaluate the acquired meanings by comparing the model's rankings against judgments of meaning paraphrases elicited experimentally from human subjects. By comparing the model-derived meaning paraphrases against human intuitions we are able to explore: (a) whether plausible meanings are ranked higher than implausible ones; (b) whether the model can be used to derive the argument preferences for a given adjective, i.e., whether the adjective is biased towards a subject or object interpretation or whether it is equi-biased; (c) whether there is a linear relationship between the model-derived likelihood of a given meaning and its perceived plausibility, using correlation analysis.

In the following sections we describe our method for assembling the set of experimental materials and eliciting judgments for model-derived adjective-noun interpretations. Section 5.4.2 reports the results of comparing human judgments to model-derived meanings, whereas Section 5.4.3 offers some discussion and concluding remarks.

5.4.1.1. Subjects

Sixty-five native speakers of English participated in the experiment. The subjects were recruited over the Internet by postings to relevant newsgroups and mailing lists. Participation was voluntary and unpaid. Subjects had to be linguistically naive, i.e., neither linguists nor students of linguistics were allowed to participate.

The data of one subject were eliminated after inspection of his response times showed that he had not completed the experiment in a realistic time frame (average response time $<$ 1000ms). The data of four subjects were excluded because they were non-native speakers of English.

This left 60 subjects for analysis. Of these, 54 subjects were right-handed, six left-handed; 22 subjects were female, 38 male. The age of the subjects ranged from 18 to 54 years, the mean was 27.4 years.

5.4.1.2. Materials and Design

We chose nine adjectives according to a set of minimal criteria and paired each adjective with 10 nouns randomly selected from the BNC. We chose the adjectives as follows: we first compiled a list of all the polysemous adjectives mentioned in the lexical semantics literature (Pustejovsky 1995; Vendler 1968). From these we randomly sampled nine adjectives (*difficult, easy, fast, good, hard, right, safe, slow*, and *wrong*). These adjectives had to be unambiguous with respect to their part of speech: each adjective was unambiguously tagged as "adjective" 98.6% of the time, measured as the number of different part-of-speech tags assigned to the word in the BNC. The nine selected adjectives ranged in BNC frequency from 80 to 1,245 per million.

We identified adjective-noun pairs using Gsearch (see Section 2.3.1 in Chapter 2 for details). Gsearch was run on a lemmatized version of the BNC so as to compile a comprehensive corpus count of all nouns occurring in a modifier-head relationship with each of the nine adjectives. From the syntactic analysis provided by the parser we extracted a table containing the adjective and the head of the noun phrase following it. In the case of compound nouns, we only included sequences of two nouns, and considered the rightmost occurring noun as the head. From the retrieved adjective-noun pairs, we removed all pairs with BNC frequency of one, as we wanted to reduce the risk of paraphrase ratings being influenced by adjective-noun combinations unfamiliar to the subjects. Furthermore, we excluded pairs with deverbal nouns (i.e., nouns derived from a verb) such as *fast programmer* since an interpretation can be easily arrived at for these pairs by mapping the deverbal noun to its corresponding verb. A list of deverbal nouns was obtained from two dictionaries, CELEX (Burnage 1990) and NOMLEX (Macleod et al. 1998, see Section 7.2.1.2 for details).

We used the model outlined in Section 5.2 to derive meanings for the 90 adjective-noun combinations. We employed no threshold on the frequencies $f(v,a)$ and $f(rel,v,n)$. As

Table 5.6: Randomly selected example stimuli with log-transformed probabilities derived by the model

Adjective-noun	Probability Band					
	High		Medium		Low	
difficult customer	satisfy	−20.27	help	−22.20	drive	−22.64
easy food	cook	−18.94	introduce	−21.95	finish	−23.15
fast pig	catch	−23.98	stop	−24.30	use	−25.66
good postcard	send	−20.17	draw	−22.71	look at	−23.34
hard number	remember	−20.30	use	−21.15	create	−22.69
right school	apply to	−19.92	complain to	−21.48	reach	−22.90
safe drug	release	−22.24	try	−23.38	start	−25.56
slow child	adopt	−19.90	find	−22.50	forget	−22.79
wrong colour	use	−21.78	look for	−22.78	look at	−24.89

in Experiment 7 the frequency $f(v,a)$ was obtained for verb-adverb modification by mapping the adjective to its corresponding adverb. For the adjectives *difficult, easy, fast,* and *good* the mapping was the same as in Experiment 7. Furthermore, the adjective *hard* was mapped to the adverb *hard,* the adjective *right* to *rightly* and *right, safe* to *safely* and *safe, slow* to *slowly* and *slow* and *wrong* to *wrongly* and *wrong.* We estimated the probability $P(a,n,rel,v)$ for each adjective-noun pair by varying both the terms v and *rel*. In other words, for each adjective-noun combination we derived both subject-related and object-related paraphrases.

In order to generate stimuli covering a wide range of model-derived paraphrases corresponding to different degrees of likelihood, for each adjective-noun combination we divided the set of the derived meanings into three "probability bands" (High, Medium, and Low) of equal size and randomly chose one interpretation from each band. The division ensured that the experimental stimuli represented the model's behavior for likely and unlikely paraphrases and enabled us to test the hypothesis that likely paraphrases correspond to high ratings and unlikely paraphrases correspond to low ratings. We performed separate divisions for object-related and subject-related paraphrases resulting in a total of six interpretations for each adjective-noun combination, as we wanted to determine whether there are differences in the model's predictions with respect to the argument function (i.e., object or subject) and also because we wanted to compare experimentally-derived adjective biases against model-derived biases. Example stimuli (with object-related interpretations only) are shown in Table 5.6 for each of the nine adjectives.

Our experimental design consisted of the factors adjective-noun pair (*Pair*), grammatical function (*Func*) and probability band (*Band*). The factor *Pair* included 90 adjective-noun combinations. The factor *Func* had two levels, subject and object, whereas the factor *Band* had three levels, High, Medium, and Low. This yielded a total of $Pair \times Func \times Band = 90 \times 2 \times 3 = 540$ stimuli. The number of the stimuli was too large for subjects to judge in

one experimental session. We limited the size of the design by selecting a total of 270 stimuli according to the following criteria: our initial design created two sets of stimuli, 270 subject-related stimuli and 270 object-related stimuli. For each set of stimuli (i.e., object- and subject-related) we randomly selected five nouns for each of the nine adjectives together with their corresponding interpretations in the three probability bands (High, Medium, Low). This yielded a total of $Pair \times Func \times Band = 45 \times 2 \times 3 = 270$ stimuli. This way, stimuli were created for each adjective in both subject-related and object-related interpretations.

We administered the 270 stimuli to two separate subject groups. Each group saw 135 stimuli consisting of interpretations for all adjective-noun pairs. For the first group five adjectives were represented by object-related meanings only (*difficult, easy, good, hard, slow*); these adjectives were presented to the second group with subject-related interpretations only. Correspondingly, for the first group four adjectives were represented by subject-related meanings only (*safe, right, wrong, fast*); the second group saw these adjectives with object-related interpretations.

Each experimental item consisted of an adjective-noun pair and a sentence paraphrasing its meaning. Paraphrases were created by the experimenter by converting the model's output to a simple phrase, usually a noun modified by a relative clause. A native speaker of English was asked to confirm that the paraphrases were syntactically well-formed (items found syntactically odd were modified and re-tested). The list of the experimental items is given in Appendix C.

5.4.1.3. Procedure

The experimental paradigm was Magnitude Estimation (ME), a technique standardly used in psychophysics to measure judgments of sensory stimuli (Stevens 1975), which Bard et al. (1996), Cowart (1997), and Keller (2000) have applied to the elicitation of linguistic judgments. ME has been shown to provide fine-grained measurements of linguistic acceptability which are robust enough to yield statistically significant results, while being highly replicable both within and across speakers (see Section 2.5.2 for further details on ME).

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

In the present experiment, each subject took part in an experimental session that lasted approximately 20 minutes; this consisted of a training phase, a practice phase and a test phase. The experiment was self-paced, and response times were recorded to allow the data to be screened for anomalies. The experiment was conducted remotely over the Internet. Subjects accessed the experiment using their web browser, which established an Internet connection

to the experimental server running WebExp 2.1 (Keller, Corley, Corley, Konieczny, and Todirascu 1998), an interactive software package for administering web-based psychological experiments. (For a detailed discussion of WebExp and the validity of web-based psycholinguistic data, see Corley, Keller, and Scheepers 2000 and Keller and Alexopoulou 2001.)

Instructions. Before participating in the actual experiment, subjects were presented with a set of instructions. The instructions explained the concept of numerical magnitude estimation of line length. Subjects were instructed to make estimates of line length relative to the first line they would see, the reference line. Subjects were told to give the reference line an arbitrary number, and then assign a number to each following line so that it represented how long the line was in proportion to the reference line. Several example lines and corresponding numerical estimates were provided to illustrate the concept of proportionality.

Then subjects were instructed to judge how well a sentence paraphrases an adjective-noun combination in the same way as line the length. Examples of plausible and implausible paraphrases were provided, together with examples of numerical estimates.

Subjects were told that they could use any range of positive numbers for their judgments, including decimals. It was stressed that there was no upper or lower limit to the numbers that could be used (exceptions being zero or negative numbers). Subjects were urged to use a wide range of numbers and to distinguish as many degrees of paraphrase plausibility as possible. It was also emphasized that there were no "correct" answers, and that subjects should base their judgments on first impressions, not spending too much time to think about any one paraphrase. The experimental instructions are given in Appendix B.

Demographic Questionnaire. After the instructions, a short demographic questionnaire was administered. The questionnaire included name, email address, age, sex, handedness, academic subject or occupation, and language region. Handedness was defined as "the hand you prefer to use for writing", while language region was defined as "the place (town, federal state, country) where you learned your first language". The results of the questionnaire are reported in Section 5.4.1.1.

Training Phase. The training phase was meant to familiarize subjects with the concept of numeric magnitude estimation using line lengths. Items were presented as horizontal lines, centered in the window of the subject's web browser. After viewing an item, the subject had to provide a numerical judgment over the computer keyboard. After pressing Return, the current item disappeared and the next item was displayed. There was no possibility to revisit previous items or change responses once Return had been pressed. No time limit was set for either the item presentation or for the response, although response times were recorded to allow inspection of the data.

Subjects first judged the modulus item, and then all the items in the training set. The modulus was the same for all subjects, and it remained on the screen all the time to facilitate

Table 5.7: Descriptive statistics for log-transformed model-derived probabilities

Rank	Mean	StdDev	StdEr	Min	Max
High	−20.49	1.71	.18	−23.99	−15.93
Medium	−22.62	.99	.10	−25.24	−20.24
Low	−23.91	.86	.18	−25.85	−22.46

comparison. Items were presented in random order, with a new randomization being generated for each subject.

The training set contained six horizontal lines. The range of the smallest to largest item was 1:10. The items were distributed evenly over this range, with the largest item covering the maximal window width of the web browser. A modulus item in the middle of the range was provided.

Practice Phase. This phase allowed subjects to practice magnitude estimation of adjective-noun paraphrases. Presentation and response procedure was the same as in the training phase, with linguistic stimuli being displayed instead of lines. Each subject judged the whole set of practice items, again in random order.

The practice set consisted of eight paraphrase sentences that were representative of the test materials (see Section 5.4.1.2 for details about the construction of experimental stimuli). The paraphrases were based on the three probability bands and illustrated subject- and object-related interpretations. A modulus item in the middle of the range was provided.

Experimental Phase. Presentation and response procedure in the experimental phase were the same as in the practice phase. Each subject group saw 135 experimental stimuli (i.e., adjective-noun pairs and their paraphrases). As in the practice phase, the paraphrases were representative of the three probability bands (i.e., High, Medium, Low) and the two grammatical functions (i.e., object, subject). A modulus item in the middle of the range was provided (see Appendix C). The modulus was the same for all subjects and remained on the screen all the time. Subjects were assigned to subject groups at random, and a random stimulus order was generated for each subject (for the complete list of experimental stimuli see Appendix C).

5.4.2. Results

The data were first normalized by dividing each numerical judgment by the modulus value that the subject had assigned to the reference sentence. This operation creates a common scale for all subjects. Then the data were transformed by taking the decadic logarithm. This transformation ensures that the judgments are normally distributed and is standard practice for magnitude estimation data (Bard et al. 1996; Lodge 1981). All analyses were conducted on the normalized, log-transformed judgments.

Table 5.8: Descriptive statistics for Experiment 8, by subjects

Rank	Mean	Std Dev	StdEr	Min	Max
High	−.0005	.2974	.0384	−.68	.49
Medium	−.1754	.3284	.0424	−.70	.31
Low	−.2298	.3279	.0423	−.68	.37

We performed an analysis of variance (ANOVA) to determine whether there is a relation between the paraphrases derived by the model and their perceived likelihood. In particular, we tested the hypothesis that meaning paraphrases assigned high probabilities by the model are perceived as better paraphrases by the subjects and correspondingly that meaning paraphrases with low probabilities are perceived as worse paraphrases. The descriptive statistics for the model-derived probabilities are shown in Table 5.7. The ANOVA revealed that the Probability Band effect was significant, in both by-subjects and by-items analyses: $F_1(2,118) = 101.46$, $p < .01$; $F_2(2,88) = 29.07$, $p < .01$. The geometric mean of the ratings in the High band was $-.0005$, compared to Medium items at $-.1754$ and Low items at $-.2298$ (see Table 5.8). Post-hoc Tukey tests indicated that the differences between all pairs of conditions were significant at $\alpha = .01$ in the by-subjects analysis. The difference between High and Medium items as well as High and Low items was significant at $\alpha = .01$ in the by-items analysis, whereas the difference between Medium and Low items did not reach significance. These results show that meaning paraphrases derived by the model correspond to human intuitions: paraphrases assigned high probabilities by the model are perceived as better than paraphrases that are assigned low probabilities.

We further explored the linear relationship between the subjects' rankings and the corpus-based model, using correlation analysis. The elicited judgments were compared with the interpretation probabilities which were obtained from the model described in Section 5.2 to examine the extent to which the proposed interpretations correlate with human intuitions. A comparison between our model and the human judgments yielded a Pearson correlation coefficient of .40 ($p < .01$, $N = 270$). Figure 5.1 plots the relationship between judgments and model probabilities. This verifies the Probability Band effect discovered by the ANOVA, in an analysis which compares the individual interpretation likelihood for each item with elicited interpretation preferences, instead of collapsing all the items in three equivalence classes (i.e., High, Medium, Low).

In order to evaluate whether the grammatical function has any effect on the relationship between the model-derived meaning paraphrases and the human judgments, we split the items into those that received a subject interpretation and those that received an object interpretation. A comparison between our model and the human judgments yielded a correlation of $r = .53$ ($p < .01$, $N = 135$) for object-related items and a correlation of $r = .21$ ($p < .05$, $N = 135$) for subject-related items. Note that a weaker correlation is obtained for subject-related inter-

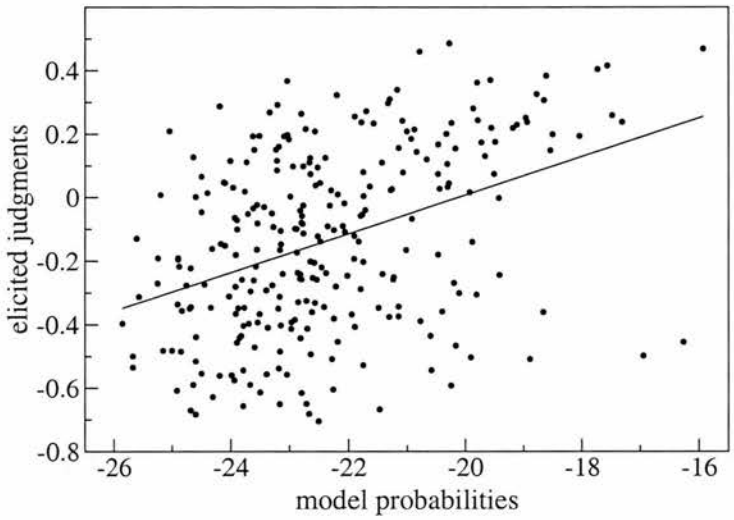

Figure 5.1: Correlation of elicited judgments and model-derived probabilities

pretations. One explanation for that could be the parser's performance, i.e., the parser is better at extracting verb-object tuples than verb-subject tuples. Another hypothesis (which we test below) is that most adjectives included in the experimental stimuli have an object-bias, and therefore subject-related interpretations are generally less preferred than object-related ones.

An important question is how well humans agree in their paraphrase judgments for adjective-noun combinations. Inter-subject agreement gives an upper bound for the task and allows us to interpret how well the model is doing in relation to humans. For each subject group we performed correlations on the elicited judgments using a method similar to leave-one-out cross-validation (Weiss and Kulikowski 1991). We divided the set of the subjects' responses with size m into a set of size $m-1$ (i.e., the response data of all but one subject) and a set of size one (i.e., the response data of a single subject). We then correlated the mean ratings of the former set with the ratings of the later. This was repeated m times. Since each group had 30 subjects we performed 30 correlation analyses and report their mean. For the first group, the average inter-subject agreement was .67 (Min = .03, Max = .82, StdDev = .14), and for the second group .65 (Min = .05, Max = .82, StdDev = .14). This means that our model performs satisfactorily given that humans do not perfectly agree in their judgments (recall that comparison between model probabilities and human judgments yielded a correlation coefficient of .40).

In sum, the correlation analysis supports the claim that adjective-noun paraphrases with high probability are judged more plausible than pairs with low probability. It also suggests that the meaning preference ordering produced by the model is intuitively correct since subjects' perception of likely and unlikely meanings correlates with the probabilities assigned by the model.

The elicited judgments can be further used to derive the grammatical function pref-

Table 5.9: Log-transformed model-derived argument preferences for polysemous adjectives

Adjective	Preference	Mean	StdDev	StdEr
difficult	√ OBJ	−21.62	1.36	.04
	SUBJ	−21.80	1.34	.05
easy	√ OBJ	−21.60	1.51	.05
	SUBJ	−22.11	1.36	.06
fast	OBJ	−24.20	1.27	.13
	√ SUBJ	−23.80	1.40	.14
good	OBJ	−22.12	1.28	.06
	SUBJ	−22.27	1.10	.07
hard	√ OBJ	−21.69	1.53	.06
	SUBJ	−22.12	1.35	.06
right	√ OBJ	−21.65	1.36	.04
	SUBJ	−21.84	1.24	.04
safe	OBJ	−22.75	1.48	.10
	√ SUBJ	−22.39	1.59	.12
slow	OBJ	−22.49	1.53	.08
	SUBJ	−22.32	1.50	.07
wrong	OBJ	−23.15	1.33	.08
	SUBJ	−23.29	1.30	.08

erences (i.e., subject or object) for a given adjective. In particular, we can determine which is the preferred interpretation for individual adjectives and compare these preferences against the ones produced by our model. Argument preferences can be easily derived from the model's output by comparing subject-related and object-related paraphrases. For each adjective we gathered all the subject- and object-related interpretations derived by the model and performed an ANOVA in order to determine the significance of the Grammatical Function effect. We interpret a significant effect as bias towards a particular grammatical function. We classify an adjective as object-biased if the mean of the model-derived probabilities for the object interpretation of this particular adjective is larger than the mean for the subject interpretation; subject-biased adjectives are classified accordingly, whereas adjectives for which no effect of Grammatical Function is found are classified as equi-biased.

The effect of Grammatical Function was significant for the adjectives *difficult* ($F(1,1806) = 8.06$, $p < .01$), *easy* ($F(1,1511) = 41.16$, $p < .01$), *hard* ($F(1,1310) = 57.67$, $p < .01$), *safe* ($F(1,382) = 5.42$, $p < .05$), *right* ($F(1,2114) = 9.85$, $p < .01$), and *fast* ($F(1,92) = 4.38$, $p < .05$). The effect of Grammatical Function was not significant for the adjectives *good* ($F(1,741) = 3.95$, $p = .10$), *slow* ($F(1,759) = 5.30$, $p = .13$), and *wrong* ($F(1,593) = 1.66$, $p = .19$). The biases for these adjectives are shown in Table 5.9. The presence of the symbol √ indicates significance of the Grammatical Function effect as well as the direction of the bias.

Table 5.10: Elicited argument preferences for polysemous adjectives

Adjective	Preference	Mean	StdDev	StdEr
difficult	√ OBJ	.0745	.3753	.0685
	SUBJ	−.2870	.2777	.0507
easy	√ OBJ	.1033	.3364	.0614
	SUBJ	−.1437	.2308	.0421
fast	OBJ	−.3544	.2914	.0532
	√ SUBJ	−.1543	.4459	.0814
good	OBJ	−.0136	.3898	.0712
	SUBJ	−.1563	.2965	.0541
hard	√ OBJ	.0030	.3381	.0617
	SUBJ	−.2543	.2436	.0445
right	√ OBJ	−.0054	.2462	.0450
	SUBJ	−.2413	.4424	.0808
safe	√ OBJ	.0037	.2524	.0461
	SUBJ	−.3399	.4269	.0779
slow	OBJ	−.3030	.4797	.0876
	√ SUBJ	−.0946	.2357	.0430
wrong	√ OBJ	−.0358	.2477	.0452
	SUBJ	−.2356	.3721	.0679

Ideally, we would like to elicit argument preferences from human subjects in a similar fashion. However, since it is unpractical to experimentally elicit judgments for all paraphrases derived by the model, we will obtain argument preferences from the judgments based on the restricted set of experimental stimuli, under the assumption that they correspond to a wide range of model paraphrases (i.e., they correspond to a wide range of probabilities) and therefore they are representative of the entire set of model-derived paraphrases. This assumption is justified by the fact that items were randomly chosen from the three probability bands (i.e., High, Medium, Low). Again we consider an adjective biased if there is a significant effect of Grammatical Function. Comparison of the mean of subject-related judgments against object-related judgments determines the direction of the bias. The ANOVA indicated that the Grammatical Function effect was significant for the adjective *difficult* in both by-subjects and by-items analyses ($F_1(1,58) = 17.98$, $p < .01$; $F_2(1,4) = 53.72$, $p < .01$), and for the adjective *easy* in both by-subjects and by-items analyses ($F_1(1,58) = 10$, $p < .01$; $F_2(1,4) = 8.48$, $p = .44$). The adjectives *difficult* and *easy* are both object-biased (see Table 5.10 which shows the biases for the nine adjectives as derived from the human judgments). The adjective *good* is equi-biased (see Table 5.10), since no effect of Grammatical Function was found ($F_1(1,58) = 2.55$, $p = .12$; $F_2(1,4) = 1.01$, $p = .37$).

The effect of Grammatical Function was significant for the adjective *hard* in the by-subjects analysis only ($F_1(1,58) = 11.436$, $p < .01$; $F_2(1,4) = 2.84$, $p = .17$), whereas for the

adjective *slow* the effect was significant by subjects and marginal by items ($F_1(1,58) = 4.56$, $p < .05$; $F_2(1,4) = 6.94$, $p = .058$). The adjective *hard* is object-biased, whereas the adjective *slow* is subject-biased (see Table 5.10). For the adjective *safe* the main effect was significant by subjects and by items ($F_1(1,58) = 14.4$, $p < .0005$; $F_2(1,4) = 17.76$, $p < .05$), and for the adjective *right* the main effect was significant in both by-subjects and by-items analyses ($F_1(1,58) = 6.51$, $p < .05$; $F_2(1,4) = 15.22$, $p = .018$). This translates into an object-bias for both *right* and *safe* (see Table 5.10). The effect of Grammatical Function was significant for the adjective *wrong* only by subjects ($F_1(1,58) = 5.99$, $p = .05$; $F_2(1,4) = 4.54$, $p = .10$) and for the adjective *fast* by subjects only ($F_1(1,58) = 4.23$, $p = .05$; $F_2(1,4) = 4.43$, $p = .10$). The adjective *wrong* has an object bias, whereas *fast* has a subject bias.

We expect a correct model to assign higher probabilities to object-related interpretations and lower probabilities to subject-related interpretations for an object-biased adjective; accordingly, we expect the model to assign on average higher probabilities to subject-related interpretations for subject-biased adjectives. Comparison of the biases derived from the model with ones derived from the elicited judgments shows that the model and the humans are in agreement for all adjectives but *slow*, *wrong*, and *safe*. On the basis of human judgments *slow* has a subject bias, whereas *wrong* has an object bias (see Table 5.10). Although the model could not reproduce this result there is a tendency in the right direction (see Table 5.9).

Note that in our correlation analysis reported above the elicited judgments were compared against model-derived paraphrases without taking argument preferences into account. We would expect a correct model to produce intuitive meanings at least for the interpretation a given adjective favors. We further examined the model's behavior by performing separate correlation analyses for preferred and dispreferred biases as determined previously by the ANOVAs conducted for each adjective (see Table 5.10). Since the adjective *good* was equi-biased we included both biases (i.e., object-related and subject-related) in both correlation analyses. The comparison between our model and the human judgments yielded a Pearson correlation coefficient of .52 ($p < .01$, $N = 150$) for the preferred interpretations and a correlation of .23 ($p < .01$, $N = 150$) for the dispreferred interpretations. The correlation for the preferred interpretations is graphed in Figure 5.2. The result indicates that our model is particularly good at deriving meanings corresponding to the argument-bias for a given adjective. However, the dispreferred interpretations also correlate significantly with human judgments, which suggests that the model derives plausible interpretations even in cases where the argument bias is overridden.

5.4.3. Discussion

We have demonstrated that the meanings acquired by our probabilistic model correlate reliably with human intuitions. These meanings go beyond the examples found in the theoretical linguistics literature. The adjective-noun combinations we interpret were randomly sampled from

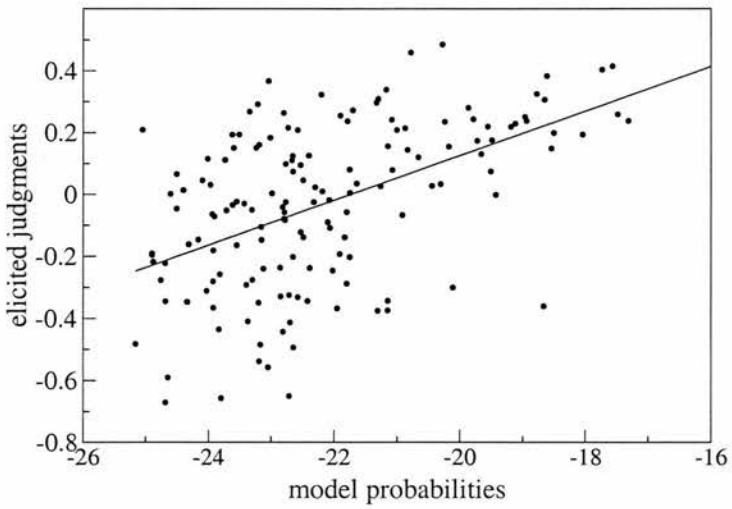

Figure 5.2: Correlation between model and human judgments for preferred argument interpretations

a large balanced corpus providing a rich inventory for their meanings. Our model does not only acquire clusters of meanings (following Vendler's 1968 insight) but furthermore can be used to obtain a tripartite distinction of adjectives depending on the type of paraphrase they prefer: subject-biased adjectives tend to modify nouns which act as subjects of the paraphrasing verb, object-biased adjectives tend to modify nouns which act as objects of the paraphrasing verb, whereas equi-biased adjectives display no preference for either argument role.

The interpretation biases generated by our model seem to correspond to human intuitions about the interpretation of polysemous adjectives. This is an important result given the simplifying assumptions underlying our probabilistic model. We have shown that the model has three defining features: (a) it is able to derive intuitive meanings for adjective-noun combinations, (b) it models the context dependency of polysemous adjectives (e.g., different meanings are predicted for *good* when it modifies *cook* and *soup*), and (c) it accurately models the argument bias of a given adjective. To address issue (c) the experimental design included both subject- and object-related interpretations for all nine adjectives. A comparison between the argument preferences produced by the model and human intuitions revealed that most adjectives (six out of nine) display a preference for an object interpretation (see Table 5.10), two adjectives are subject-biased (i.e., *fast*, *slow*) and one adjective is equi-biased (i.e., *good*).

Note finally that the evaluation procedure to which we subject our model is rather strict. The derived adjective-noun combinations were evaluated by subjects naive to linguistic theory. Although adjective-noun polysemy is a well researched phenomenon in the theoretical linguistics literature, the experimental approach advocated here is new to our knowledge. Despite the fact that human data is noisy as evidenced by the fairly low inter-subject agreement (.67 for the first group and .65 for the second group) we obtain reliable correlations between

elicited judgments and model predictions.

The probabilistic model described in Section 5.2 explicitly takes adjective/adverb and verb co-occurrences into account. However, one could derive meanings for polysemous adjective-noun combinations by solely concentrating on verb-noun relations, ignoring thus the adjective/adverb and verb dependencies. For example, in order to interpret the combination *easy problem* we could simply take into account the types of activities which are related with problems (i.e., solving them, giving them, etc.). This simplification is consistent with Pustejovsky's (1995) claim that polysemous adjectives like *easy* are predicates, modifying some aspect of the head noun and more specifically the events associated with the noun. A "naive baseline" model would be one which simply takes into account the number of times the noun in the adjective-noun pair acts as the subject or object of a given verb, ignoring the adjective completely. This raises the question of how well would such a naive model, which takes only verb-argument relations into account, perform at deriving meaning paraphrases for polysemous adjective-noun combinations.

In the following section we present a naive model of adjective-noun polysemy. We compare the model's predictions against the elicited judgments. Using correlation analysis we attempt to determine whether the naive model can provide an intuitively plausible ranking of meanings (i.e., whether perceived likely/unlikely meanings are given high/low probabilities). We further compare the naive model to our initial model (see Section 5.2) and discuss their differences.

5.5. Experiment 9: Comparison against Naive Baseline

In this section we present a naive model which does not take adjective or adverb and verb modification into account. The model relies solely on verb-object and verb-subject tuples extracted from the corpus. In Section 5.5.3 we evaluate the model's performance through comparisons to human judgments and the interpretation preferences derived by our initial model (see Section 5.2).

5.5.1. Naive Baseline Model

Given an adjective-noun combination we are interested in finding the events most closely associated with the noun modified by the adjective. In other words we are interested in the verbs whose object or subject is the noun appearing in the adjective-noun combination. This can be simply expressed as $P(v|rel,n)$, the conditional probability of a verb v given an argument-noun relation rel,n:

$$P(v|rel,n) = \frac{f(v,rel,n)}{f(rel,n)} \tag{5.24}$$

The model in 5.24 assumes that the meaning of an adjective-noun combination is independent of the adjective in question. Consider for example the adjective-noun pair *fast plane*. We need to find the verbs v and the argument relation *rel* that maximize the probability $P(v|rel,plane)$. Intuitively speaking, the model in (5.24) takes into account only the verbs that are associated with the noun modified by the adjective. In the case of *fast plane* the verb that is most frequently associated with planes is *fly* (see Table 5.2 in Section 5.2.1). Note that this model will come up with the same probabilities for *fast plane* and *wrong plane* since it does not take the identity of the modifying adjective into account. We estimated the frequencies $f(v,rel,n)$ and $f(rel,n)$ from verb-object and verb-subject tuples extracted from the BNC using Cass (Abney 1996) (see Section 5.2.2 for details on the extraction and filtering of the argument tuples).

5.5.2. Method

Using the naive model we calculated the meaning probability for each of the 270 stimuli included in Experiment 8. Through correlation analysis we explored the linear relationship between the elicited judgments and the naive baseline model. We further directly compared the two models, our initial, linguistically more informed model, and the naive baseline. We report our results in the following section.

5.5.3. Results

Using correlation analysis we explored which model performs better at deriving meaning paraphrases for adjective-noun combinations. A comparison between the naive model's probabilities and the human judgments yielded a Pearson correlation coefficient of .25 ($p < .01$, $N = 270$). Recall that we obtained a correlation of .40 ($p < .01$, $N = 270$) when comparing our original model to the human judgments. Not surprisingly the two models are intercorrelated ($r = .38$, $p < .01$, $N = 270$). These correlations are shown in Table 5.11 where 'Model' refers to our initial model and 'Baseline' refers to the naive baseline. An important question is whether the difference between the two correlation coefficients ($r = .40$ and $r = .25$) is due to chance. Comparison of the two correlation coefficients revealed that their difference was significant ($t(267) = 2.42$, $p < .01$). This means that our original model (see Section 5.2) performs reliably better than a naive baseline at deriving interpretations for polysemous adjective-noun combinations.

We further compared the naive baseline model and the human judgments separately for subject-related and object-related items. The comparison yielded a correlation of $r = .29$ ($p < .01$, $N = 135$) for object interpretations. Recall that our original model yielded a correlation coefficient of .53 (see Table 5.12). The two correlation coefficients were significantly different ($t(132) = 3.03$, $p < .01$). No correlation was found for the naive model when com-

Table 5.11: Correlation matrix for human judgments and the two corpus-based models

	Judgments	Model
Model	.40**	
Baseline	.25**	.38**
*$p < .05$ (2-tailed)		**$p < .01$ (2-tailed)

Table 5.12: Correlation matrices for human judgments and the two corpus-based models

OBJECT		
	Judgments	Model
Model	.53**	
Baseline	.29**	.42**
*$p < .05$ (2-tailed)		**$p < .01$ (2-tailed)

SUBJECT		
	Judgments	Model
Model	.20*	
Baseline	.09	.37**
*$p < .05$ (2-tailed)		**$p < .01$ (2-tailed)

pared against elicited subject interpretations ($r = .09$, $p = .28$, $N = 135$, see Table 5.12).

5.5.4. Discussion

We have demonstrated that a naive baseline model which interprets adjective-noun combinations by focusing solely on the events associated with the noun is outperformed by a more detailed model which not only considers verb-argument relations but also adjective-verb and adverb-verb dependencies. Although the events associated with the different nouns are crucially important for the meaning of polysemous adjective-noun combinations, it seems that more detailed linguistic knowledge is needed in order to produce intuitively plausible interpretations. This is by no means surprising. To give a simple example consider the adjective-noun pair *fast horse*. There is a variety of events associated with the noun *horse*, yet only a subset of those are likely to occur fast. The three most likely interpretations for *fast horse* according to the naive model are "a horse that needs something fast", "a horse that gets something fast", "horse that does something fast". A model which uses information about verb-adjective or verb-adverb dependencies (see Section 5.2) provides a more plausible ranking: a *fast horse* is "a horse that runs, learns, or goes fast". A similar situation arises when one considers the pair *careful scientist*. According to the naive model a *careful scientist* is more likely to "believe, say, or make something carefully". However, none of these events are particularly associated with the adjective *careful*.

5.6. General Discussion

In this chapter we focused on polysemous adjective-noun combinations. We showed how adjectival meanings can be acquired from a large corpus and provided a probabilistic model which derives a preference ordering on the set of possible interpretations. In contrast to the study

presented in Chapter 4 (where the meanings of verbs were provided by Levin's 1993 linguistic classification) the meanings for polysemous adjectives were derived solely from the corpus using a surface cueing approach. The probabilistic model reflects linguistic observations about the nature of polysemous adjectives: it predicts the context dependency effect (i.e., the meaning of the adjective varies with respect to the noun it modifies) and is faithful to Vendler's (1968) claim that polysemous adjectives are usually interpreted by a cluster of meanings instead of a single meaning.

Furthermore, the proposed model can be viewed as complementary to linguistic theory: it automatically derives a ranking of meanings, thus distinguishing likely from unlikely interpretations. Even if linguistic theory was able to enumerate all possible interpretations for a given adjective (note that in the case of polysemous adjectives we would have to take into account all nouns or noun classes the adjective could possibly modify) it has no means to indicate which ones are likely and which ones are not. Our model fares well on both tasks. It recasts the problem of adjective-noun polysemy in a probabilistic framework and approximates meaning by taking into account the relation of a verb to its argument (i.e., the noun the adjective is in construction with) together with the relation of the same verb and the adjective (or its corresponding adverb), deriving thus a large number of interpretations not readily available from linguistic introspection. The information acquired from the corpus can be also used to quantify the argument preferences of a given adjective. These are only implicit in the lexical semantics literature where certain adjectives are exclusively given a verb-subject or verb-object interpretation (see the adjectives *fast* and *difficult* in Table 5.4). We have demonstrated that we can empirically derive argument biases for a given adjective that correspond to human intuitions.

Our model is ignorant about the potential different meanings of the noun in the adjective-noun pair. For example, the combination *fast plane* may be a fast aircraft, or a fast tool, or a fast geometrical plane. Our model derives meanings related to all three senses of the noun *plane*. For example, a *fast plane* is not only "a plane (i.e., an aircraft) which flies, lands, or travels quickly", but also "a plane (i.e., a surface) which transposes or rotates quickly" and "a plane (i.e., a tool) which smoothes something quickly". However, more paraphrases are derived for the aircraft sense of plane; these paraphrases also receive a higher ranking. This is not surprising since the number of verbs related with the aircraft sense of plane are more frequent than the verbs related with the other two senses. Note also that *fast* is more likely to be related with motion verbs than verbs related to the events denoted by *plane* in the sense of "surface" or "tool". There are also cases where a model-derived paraphrase does not provide disambiguation clues with respect to the meaning of the noun. Consider the adjective-noun combination *fast game* from Section 5.3. The model comes up with the paraphrases "game that runs fast" or "game that goes fast". Both paraphrases may well refer to either the "contest", "activity", or "prey" sense of *game*.

Our results provide further support for the surface cueing approach. In Chapter 3 we

used surface cues as indicators about meaning. Similarly in this chapter we derive the meanings of adjective-noun combinations by taking into account co-occurrence frequencies acquired through shallow syntactic processing of the corpus. Recall that in Chapter 3 we used the acquired frequencies to quantify and augment Levin's (1993) generalizations about alternating verbs. Here we examined the empirical validity of Pustejovsky's (1995) and Vendler's (1968) claims about context-sensitive adjectives. In Chapter 4 we showed how a probabilistic model which uses Levin's taxonomy as an inventory of verb meanings can make use of co-occurrence frequencies to derive a ranking of interpretations for polysemous verbs. Although a similar approach is taken in this chapter (our probabilistic model provides a preference ordering on the set of acquired interpretations for polysemous adjective-noun combinations) adjective meanings are derived directly from the corpus without recourse to a predefined taxonomy.

The acquired adjective-noun meanings could be potentially useful for a variety of NLP tasks. One obvious application is Natural Language Generation. The acquisition task can be cast in terms of finding the corresponding paraphrase for a given adjective-noun combination. For example, a generator that has knowledge of the fact that *fast plane* corresponds to "a plane that flies fast" can exploit this information either to render the text shorter (in cases where the input representation is a sentence) or longer (in cases where the input representation is an adjective-noun pair). Information retrieval is another application that easily comes to mind. Consider a search engine faced with the query *fast plane*. Presumably one would not like to obtain information about planes in general or about planes that go down or burn fast but rather about planes that fly or travel fast. So knowledge about the most likely interpretations of *fast plane* could help rank relevant documents before non-relevant ones or restrict the number of retrieved documents.

5.7. Related Work

Previous corpus-based work relating to adjectives has focused on two directions: the automatic classification of adjectives in terms of their semantic features (e.g., gradable, marked, positive, negative, see Section 5.1) and the disambiguation of their senses. Work in classification has concentrated on exploring the contribution of several linguistic indicators (using machine learning) for determining the semantic behavior of a given adjective. The approach aims at determining what is the most likely class for an adjective rather than determining what is its specific class in a given context. The word sense disambiguation approach aims at determining the meaning of the adjective within its surrounding context. Our work is not a classification task, we aim at discovering meanings for polysemous adjective-noun combinations, using, however, linguistic indicators. Our model provides a ranking on the set of possible meanings, ignoring the context surrounding a given adjective-noun pair. Our adjective-noun paraphrases can be thought of as input to a sense disambiguation process which would have to choose among

them. In what follows we review work on classification and word sense disambiguation and compare it to our own work.

Hatzivassiloglou and McKeown (1995b) present an empirical method which discovers semantically related adjectives. Their approach makes use of simple co-occurrence frequencies (adjective-noun and adjective-adjective pairs) to measure the similarity between adjectives without recourse to linguistic information other than the one present in the corpus. The derived groups of semantically related adjectives can be further filtered so as to distinguish gradable from non-gradable adjectives. Hatzivassiloglou and Wiebe (2000) propose a log-linear statistical model that classifies adjectives in terms of their gradability (i.e., gradable or non-gradable). The model achieves a high precision (87.97%) by simply taking into account the number of times an adjective has been observed in the context of a degree modifier (e.g., *very*).

Hatzivassiloglou and McKeown (1995a) develop a method for selecting the semantically unmarked term out of a pair of antonymous adjectives (see Section 5.1 for details on markedness). The approach exploits several linguistic diagnostics for markedness such as text frequency (unmarked terms are more frequent than marked ones) and morphological complexity (unmarked terms are morphologically simpler). Hatzivassiloglou and McKeown's results show that the best predictor for markedness is frequency achieving an accuracy of 80.64%. A similar approach is put forward in Hatzivassiloglou and McKeown (1997) in order to automatically identify the semantic orientation of adjectives (see Section 5.1). The approach correlates linguistic indicators with semantic orientation. For example, in most cases coordinated adjectives are of the same orientation (e.g., *fair and legitimate*); when the connective is *but*, the adjectives are usually of different orientation. Hatzivassiloglou and McKeown present a log-linear regression model which relies solely on information present in conjunctions of adjectives (extracted from a corpus) and achieves a precision of 82% at determining if two conjoined adjectives are of the same or different orientation. A clustering algorithm is used to separate the adjectives in two subsets of different orientation (i.e., positive or negative). The approach receives a 92% accuracy on the classification task.

The approach put forward by Justeson and Katz (1995b) uses nouns as indicators for discriminating among the senses of adjectives that modify them. Justeson and Katz observe that some nouns in adjective-noun combinations are strongly associated with specific adjectival senses. For example, when the adjective *old* modifies the noun *man*, it is typically used in the sense "aged", whereas when the same adjective modifies the noun *house* it is used in the sense "not new". Justeson and Katz discover which nouns are reliable indicators of a particular adjective sense by looking at antonyms modifying the same noun in corpus sentences. For example, in a sentence where *old* and *young* modify the noun *man* it is safe to assume that *old* is interpretable as "not young", whereas in a sentence where *old* and *new* co-occur as modifiers of the noun *house*, *old* is interpretable as "not new". Justeson and Katz's study focuses on five ambiguous adjectives (*hard, light, old, right*, and *short*) with two antonym-related senses

(e.g., hard relates to "not easy" and "not soft") and shows that adjectives can be disambiguated with very high precision (97%) on the basis of their nouns.

Chao and Dyer (2000) propose a method for the disambiguation of polysemous adjectives which exploits WordNet's taxonomic information. More specifically, Chao and Dyer introduce a probabilistic model which estimates the likelihood of each adjective sense given the semantic features of the noun it modifies. WordNet's inventory of senses is used both for adjectives and nouns, while the model's parameters are estimated by submitting queries (e.g., *great hurricane*) to the Altavista search engine and extrapolating from the number of returned documents the frequency of the adjective-noun pair (see Mihalcea and Moldovan 1998 for details of this technique). Manually disambiguated adjective-noun combinations are represented in terms of Bayesian belief networks (encoding the distribution of an adjective sense over the semantic features of the noun) which are in turn used to disambiguate unseen adjective-noun combinations. The method achieves a precision of 81.4%. Chao and Dyer's approach is conceptually similar to Justeson and Katz's (1995b) work. The main assumption underlying both proposals is that the noun modified by the adjective in question plays a key role in its disambiguation.

Our approach focuses on a novel task, the interpretation of systematically polysemous adjectives (i.e., adjectives whose meanings are not fixed but vary with respect to the noun they modify). In contrast to Chao and Dyer (2000) and Justeson and Katz (1995b) we do not employ a static predefined inventory of adjective senses (e.g., WordNet or antonymic relations). Instead, we derive the meanings of polysemous adjective-noun combinations dynamically from a large balanced corpus. Similarly to Hatzivassiloglou and McKeown (1995a, 1997), Hatzivassiloglou and Wiebe (2000), and Chao and Dyer (2000) our approach is probabilistic: the acquired meanings are ranked in terms of their likelihood in the corpus. We estimate the parameters of our model straightforwardly by approximating the meaning of an adjective-noun pair to a verb which is modified by the adjective (or its corresponding adverb) and whose subject or object is the noun the adjective is in construction with. In contrast to Chao and Dyer, our model makes minimal assumptions about how the meaning of a word is represented and combined with the meaning of other words.

Although our approach patterns with Hatzivassiloglou and McKeown (1995a, 1997), and Hatzivassiloglou and Wiebe (2000) in that it exploits linguistic diagnostics (e.g., verb-adverb dependencies, verb-argument relations) for deriving the interpretations of polysemous adjectives, it goes beyond finding correspondences between linguistic features and corpus data. Experiments 7 and 8 examine the empirical basis of theoretical generalizations about the behavior of context sensitive adjectives (Pustejovsky 1995; Vendler 1968). We showed that the meanings derived by our model not only correspond to the interpretations discussed in the lexical semantics literature (see Section 5.3) but also to human intuitions (see Section 5.4). We were further able to quantify implicit assumptions in the lexical semantics literature such the interpretation bias (i.e., subject, object, or none) of a given adjective.

5.8. Summary

In this chapter we investigated polysemous adjectives whose meaning varies depending on the nouns they modify. We acquired the meanings of these adjectives from a large corpus and proposed a probabilistic model which provides a ranking on the set of possible interpretations. We identified lexical semantic information automatically by exploiting the consistent correspondences between surface syntactic cues and lexical meaning.

We evaluated our results against paraphrase judgments elicited experimentally from subjects naive to linguistic theory and showed that the model's ranking of meanings correlates reliably with human intuitions: meanings that were found highly probable by the model were also rated as plausible by the subjects. More specifically, comparison between our model and human judgments yields a reliable correlation of .40 when the upper bound for the task (i.e., inter-subject agreement) is approximately .65. Furthermore, our model performs reliably better than a naive baseline model, which only achieves a correlation of .25.

In the next chapter we turn to noun-noun modification. We focus on the acquisition of noun-noun compounds, a very productive lexical phenomenon and further examine whether the surface cueing approach can be used in order to discover not only frequent but also rare compounds. The latter are not exceptional or idiosyncratic but account for more than half of the noun-noun sequences found in the corpus.

Chapter 6

Compound Nouns

The present chapter focuses on the acquisition of compound nouns in domain independent wide-coverage text. In Chapter 3 we presented a series of experiments on the acquisition of alternating verbs using a combination of linguistic insights (i.e., Levin's 1993 generalizations about the correspondence between syntax and meaning) and surface syntactic and semantic cues. In the present chapter we pursue further the surface cueing approach by looking at compounding. We investigate the suitability of a number of statistical scores for determining whether a sequence of two nouns is a valid compound. We present several experiments which show that corpus-based measures such as the log-likelihood ratio and co-occurrence frequency can be used for the acquisition of compounds with a high degree of accuracy. However, these compounds represent only a small fraction of the candidate compounds present in the corpus, the majority of which are attested only once. We further concentrate on the acquisition of compound nouns for which very little evidence is found in the corpus and investigate how surface cues can provide useful information, even in this case. We show how evidence about established (i.e., frequent) compounds can be used to estimate features that can discriminate rare valid compounds from rare nonce terms. We use a variety of linguistic features (e.g., the likelihood of a noun as a compound modifier or a compound head, the context surrounding the candidate compound) and explore their individual and combined contribution using decision tree learning.

6.1. Introduction

The nature and properties of compounds have been studied at length in the theoretical linguistics literature. It is a well-known fact that English permits the free formation of compound nouns, the commonest type being those formed by a sequence of nouns (see the examples in (6.1)). Definitions of compound nouns in the literature employ different criteria in order to determine whether a sequence of words is a compound or not. According to Quirk et al. (1985)

any noun modified by a constituent preceding it counts as a compound noun. The definition includes not only noun sequences (see (6.1)), but also adjective-noun sequences (see (6.2)), and nouns modified by possessives (see (6.3)) or proper names (see (6.4)).

(6.1) a. bathroom
b. public-relations
c. income tax
d. income tax relief

(6.2) a. black belt
b. medical department
c. medical department associate

(6.3) a. crow's nest
b. bachelor's degree
c. bachelor's degree requirement

(6.4) a. AT & T headquarters
b. BT line
c. BT line repair

For Levi (1978) compounds are noun sequences and adjective-noun sequences with non-predicative adjectives (i.e., adjectives that do not appear in copula constructions). This definition includes examples (6.1) and (6.2b,c) (*medical* is a non-predicative adjective, **The department is medical*) but not (6.2a) (where *black* is a predicative adjective, *The belt is black*). Downing (1977: 810) employs a more restrictive definition: a compound is the concatenation of any two or more nouns functioning as a third nominal. This definition includes only the compounds in (6.1) and (6.4). Chomsky and Halle (1968) use stress as a diagnostic for compound formation: only word sequences receiving primary stress are compounds. According to this definition (6.2a) is not a compound when the stress is on *belt* (i.e., *black belt* is a belt which is black as opposed to a belt worn by one who has attained a certain degree of proficiency in judo).

Although compounds are typically binary (i.e., they are formed by two words), they can also be longer than two words (see (6.1d), (6.2c), (6.3c), and (6.4c)). The orthographic conventions for encoding compounding are varied. Compounds are commonly written as a concatenation of words (see examples in (6.1c,d)), or as single words (see (6.1a)), sometimes a hyphen is also used (see (6.1b)).

The use of noun compounds is frequent not only in technical writing and newswire text (McDonald 1982) but also in fictional prose (Leonard 1984), and spoken language (Liberman and Sproat 1992). Compounding is a very productive lexical phenomenon. Novel compounds are used as a text compression device (Marsh 1984), i.e., to pack meaning into a minimal amount of linguistic structure, as a deictic device (Downing 1977), or as a means to classify

an entity which has no specific name (Downing 1977). Novel compounds are commonly distinguished from lexicalized ones. Lexicalized compounds have a conventionalized meaning, whereas newly created compounds may be interpretable in a number of ways. As Downing (1977) observes:

> Novel compounds are coined to satisfy the speaker's need to refer to an entity which possesses no name of sufficient specificity for his classificatory or communicative purposes. (Downing 1977: 824)
>
> If a novel compound survives beyond the situation in which it is coined, it is subject to historical processes and it acquires more and more of the characteristics of a unitary lexical item. (Downing 1977: 836)

Downing (1977) and Warren (1978) point out that despite the functional differences between novel and lexicalized compounds it is difficult to rigorously distinguish between them. Downing (1977: 839) remarks that "it is not a straightforward task to decide at what point a novel compound becomes a lexicalized compound, acquiring a unitary character, surrendering to some extent its original decomposability and becoming a potential model for the creation of new compounds".

Computational investigations of compound nouns have concentrated on their automatic acquisition from corpora, syntactic disambiguation (i.e., determine the structure of compounds like *income tax relief*), and semantic interpretation (i.e., determine what is the semantic relation between *income* and *tax* in *income tax*). The acquisition of compound nouns is usually subsumed under the general discovery of terms from corpora. Terms are typically acquired by either symbolic or statistical means. Under a symbolic approach, candidate terms are extracted from the corpus using surface syntactic analysis (Bourigault 1992; Bourigault and Jacquemin 1999; Justeson and Katz 1995c; Lauer 1995) and sometimes are further submitted to experts for manual inspection. The approach typically assumes no prior terminological knowledge, although Jacquemin (1996) has proposed the detection of terminological variants in a corpus by making use of lists of existing terms.

The main assumption underlying the statistical approach to term acquisition is that lexically associated words tend to appear together more often than expected on the basis of their individual occurrence frequencies. Once candidate terms are detected in the corpus, statistical scores are used to determine which co-occurrences are valid terms. Several statistical scores have been proposed in the literature (see Daille 1996 and Manning and Schütze 1999 for overviews), the most popular being mutual information (Church and Hanks 1990) and the log-likelihood ratio (Dunning 1993).

With the exception of Lauer (1995), none of the approaches, statistical or symbolic, explicitly addresses the acquisition of compound nouns. The corpora used vary in size (see Table 6.1), in most cases they are representative of a particular domain (e.g., medicine) and

Table 6.1: Corpora used for the extraction of terms and compounds

	Corpus	Words	Domain
Church and Hanks (1990)	Associated Press (1987) Associated Press (1988)	15 M 36 M	newswire
Bourigault (1992)	Electricity Board	1.2 M	electricity
Daille (1996)	Satellite Handbook Livre bleu du CCITT	200 K 800 K	telecommunication
Lauer (1995)	Grolier's Encyclopedia	8 M	general
Jacquemin (1996)	Medic	1.56 M	medicine
Bourigault and Jacquemin (1999)	Broussais DER Menelas	40 K 230 K 110 K	medicine engineering medicine

exhibit a uniform writing style (e.g., reports, newswire text, encyclopedic text). On the one hand, this is a desirable feature since the main goal is the acquisition of terminology for a given domain. On the other hand, this means that the methodology has been developed for a particular domain and/or register and may not generalize to unrestricted text. Furthermore, by focusing strictly on the acquisition of terms one does not pay heed to the phenomenon of compounding per se, which is an extremely productive process (Downing 1977; Sparck Jones 1983).

The discovery of compound nouns is inherently different from terminological acquisition. The former typically focuses on the acquisition of noun sequences, whereas the latter concerns a wider spectrum of noun phrases containing adjectives, nouns, and prepositions. The crucial difference between compounds and technical terms relies on the processes underlying their creation. As Justeson and Katz (1995c) point out, repetition is a good diagnostic for distinguishing terminological noun phrases. The repetition criterion clearly does not apply in the case of novel compounds which can be either idiosyncratic or the result of a highly productive lexical rule. As an example consider the compounds *apple-juice seat* and *wood seat* in (6.5a) and (6.5b), respectively. The former compound can be interpreted only if we know that a glass of apple-juice has been placed in front of the seat in which the friend has been instructed to sit (Downing 1977). Although *wood seat* is attested in the BNC only once, it can be easily interpreted—even without taking context into account—since it is the result of a productive lexical rule that combines a substance (i.e., *wood*) with an artefact (i.e., *seat*) to produce an artefact made of the substance (i.e., *wood seat* is a seat made of wood).

(6.5) a. A friend of mine was once instructed to sit in the [apple-juice$_N$ seat$_N$]. (Downing 1977: 818)

b. Although no one will doubt their possibilities for elegance and robustness, sitting on a solid [wood$_N$ seat$_N$] can test the limits of comfort after quite a short time and woven seats are little better.

c. The building of new prisons and the improvement of old ones were low priorities in the aftermath of the [world$_N$ war$_N$] when scarce resources were concentrated upon houses, schools, hospitals and roads.

d. The use of the [term$_N$ shilling$_N$] derives from an 19th century system of invoicing beer according to its gravity.

Methodologies for the acquisition of compound nouns from corpora must discover not only lexicalized compounds such as *world war* (see example (6.5c)), but also novel terms such as *wood seat* resulting from the application of productive lexical rules. The acquisition of non-established compounds is particularly challenging for statistical approaches. Most of the statistical tests that have been proposed in the literature for the discovery of terms rely on the fact that candidate terms will occur frequently in the corpus (Daille 1996; Justeson and Katz 1995c) or, when hypothesis testing is applied, on the assumption that two words form a term when they co-occur more often than chance (Church and Hanks 1990; Church and Mercer 1993; Daille 1996). This means that statistical tests cannot be applied reliably for candidate compounds with co-occurrence frequency of one and cannot be used to distinguish rare but valid noun compounds (see examples (6.5a)–(6.5c)) from rare but nonce noun sequences (see the noun-noun sequence *term shilling* in sentence (6.5d)).

In the following sections we focus on the acquisition of compound nouns from the BNC, a domain independent wide-coverage corpus. We restrict our attention to compounds formed by a concatenation of nouns (see (6.1c)), ignoring for the moment other types of compound formation (see the examples in (6.2)–(6.4)). In Experiment 10 we show that a simple heuristic which looks for consecutive nouns in the corpus results in substantially lower accuracy than previously reported in the literature (Lauer 1995). Experiments 11–13 evaluate the appropriateness of several statistical scores for the acquisition of compound nouns from unrestricted text, whereas Experiment 14 examines the properties of candidate compounds which are attested in the BNC only once. Experiment 15 focuses on the acquisition of valid compounds for which very little evidence is found in the corpus.

6.2. Experiment 10: Compound Noun Extraction

6.2.1. Method

The extraction of compound nouns from a corpus has been previously addressed by Lauer (1995). He identified two word compounds in a corpus derived from the Grolier Multimedia Encyclopedia using a heuristic which simply looks for consecutive pairs of nouns. The heuristic ignores noun pairs preceded or succeeded by a noun in order to avoid identifying as two word compounds noun sequences which are part of a larger compound. Lauer did not use a part-of-speech tagged version of Grolier's Encyclopedia. Instead, he identified tokens in the corpus

Table 6.2: Noun sequences extracted from the BNC

length	2	3	4	5	6	7	8	9	> 9
tokens	2,406,588	338,349	57,098	10,354	2,059	576	209	102	241

that were members of a predefined list of 90,000 nouns which had no part-of-speech ambiguity. The heuristic, taken from Lauer (1995: 161), is shown in (6.6) below.

(6.6) $C = \{(w_2, w_3) \mid w_1\ w_2\ w_3\ w_4; w_1, w_4 \notin N; w_2, w_3 \in N\}$

Here, $w_1\ w_2\ w_3\ w_4$ denotes the occurrence of a sequence of four words in the corpus and N is the predefined set of unambiguous nouns. Although the heuristic misclassifies noun sequences which are parts of a double object construction (e.g., *The defendants were not at liberty to divulge to* [*others information*]) and nouns which are adjacent but represent parts of distinct noun phrases (e.g., *In the last few* [*years feminists*] *have begun to organize together*), Lauer (1995) reports an accuracy of 97.9% on a sample of a 1,000 noun-noun sequences.

The corpus used by Lauer (1995) contained approximately eight million words and was composed of scientific articles on subjects such as history, language and literature, geography, and art. Although the corpus is representative of several domains, the writing style is uniform across different articles and characteristic of encyclopedic text. Given the impressive accuracy of Lauer's approach and its simplicity, the question arises as to whether a comparable performance can be obtained in a larger corpus such as the BNC, which contains written and spoken text representing a variety of registers.

We used the part-of-speech tagged version of the BNC (90M words written and 10M words spoken text) to extract noun sequences of arbitrary length. Noun sequences were identified using Gsearch (Corley et al. 2001, see Section 2.3.1 for details). Gsearch was run on a lemmatized version of the BNC in order to compile a comprehensive count of all nouns occurring in a modifier-head relationship. As shown in Table 6.2 noun sequences of length two are by far the most frequent in the corpus. Tokens containing noun sequences of length two were classified as candidate compounds unless: (a) the two consecutive nouns were preceded or succeeded by a noun (e.g., *system integration manager*, *light bulb phobia*, see the heuristic in (6.6)), (b) either noun was a proper name (e.g., *London house*, *husband Michael*), and (c) either noun was a number (e.g., *£19.95 investment*, *flour 100g*). This procedure resulted in a total of 1,624,915 tokens consisting of 510,673 distinct types of candidate compounds.

6.2.2. Results

Obviously one would like to know how many of the noun-noun sequences identified by the procedure described in the previous section are valid compounds and furthermore how many are missed. In order to answer these questions we randomly selected a sample of 870 tokens from

the noun-noun sequences that were classified as compounds. Accordingly, a random sample of 800 tokens was selected from the sequences that were discarded as non-compounds, i.e., noun-noun sequences for which heuristics (a)–(c) applied (see Section 6.2.1). The noun sequences contained in the samples were manually inspected within context using the corpus concordance tool Xkwic (Christ 1995) and classified as to whether they formed a valid compound or not. The heuristics achieved an accuracy of 71.0% in identifying valid compounds (true positives out of true and false positives) and an accuracy of 98.8% in rejecting tokens containing noun sequences of length two which were not compounds (true negatives out of true and false negatives). Note that the accuracy reported here is substantially lower than the figure of 97.9% given by Lauer (1995).

An analysis of the misclassifications revealed that in some cases they were due to the absence of structural information (i.e., parsing). Consider example (6.7a), where *people* and *ideas* are the objects of the ditransitive verb *gave*, and sentences (6.7b,c) where the candidate nouns happen to be adjacent but do not form a compound. Another source of errors were tagging mistakes (see (6.8a), where the word *Alf* is tagged as a noun instead of a proper name), foreign terms (see (6.8b), where in fact *pièce d'occasion* is a compound in French), acronyms (see (6.8c)), and finally non-compositional sequences such as *pip pip* (see the sentences in (6.9)).

(6.7) a. They said it gave [people$_N$ ideas$_N$].
b. In the last five [years$_N$ evaluation$_N$] has received much attention in UK academic libraries.
c. I've answered that [way$_N$ sir$_N$] because I can't be sure.

(6.8) a. There survives a copy of a letter sent to [priest$_N$ Alf$_N$].
b. In between comes a short [pièce$_N$ d'occasion$_N$] by Brahms, written in 1853 as a surprise birthday present for the violinist Joachim.
c. Imperial Pottery and Alexandra Pottery came out joint winners in the latest bi-annual [TNT$_N$ awards$_N$].

(6.9) a. No Englishman that I have ever met said [pip$_N$ pip$_N$].
b. They had a terrific start and after that it was [rythm$_N$ rythm$_N$] all the way Oxford were beaten.

Statistical scores (e.g., mutual information, the log-likelihood ratio) have been extensively used for the acquisition of terms from domain specific corpora. We next examine whether they can be applied to the acquisition of compound nouns. More specifically, we explore whether a particular score is appropriate for compound acquisition, i.e., an appropriate score should assign high values to valid compounds and low values to non-compounds. Section 6.3 describes the statistical scores we used, and Sections 6.4–6.6 report their performance in detecting compounds.

6.3. Statistical scores

We assessed five statistical scores as quantitative indicators of the compounding relation between two nouns: the corpus co-occurrence frequency, the conditional probability of the modifier given its head noun, the informational contribution of a noun-noun pair, the log-likelihood ratio, and the mutual information.

Co-occurrence frequency. Previous work in term acquisition has shown that the exact repetition of a sequence of words in the corpus is a good indicator of its terminological status (Daille 1996; Justeson and Katz 1995c). Similarly, we examine whether the co-occurrence frequency of noun-noun pairs $f(n_1,n_2)$ provides a good cue for compound detection. Consider the two candidate compounds *world war* and *term shilling* shown in Table 6.3. Here, co-occurrence frequency (CoocF) predicts that *world war* is a far more likely compound than *term shilling*, as the former is attested in the corpus 3,707 times and the latter only once.

Conditional probability. We express the likelihood of a given noun n_1 being paired with another noun n_2 in a noun-noun sequence n_1 n_2 as the conditional probability of the modifier n_1 given the head n_2. We assume here that the head of a compound is its rightmost occurring noun (Spencer 1991). We estimate the conditional probability as shown in (6.10) below.

$$P(n_1|n_2) = \frac{f(n_1,n_2)}{f(n_2)} \tag{6.10}$$

If n_1 co-occurs only with n_2 the conditional probability will be one. Consider again the candidate compounds *world war* and *term shilling*. The probability of seeing the word *war* preceded by the word *world* is considerably higher than the probability of seeing *shilling* preceded by *term* (see Table 6.3).

Informational contribution. This statistic was introduced by Strzalkowski and Vauthey (1992) and was used to measure the relative strength of association between a head and its modifier, where the head is either a verb or a noun. Here we adapt the informational contribution measure to account for noun-noun modification only. The informational contribution[1] is based on the conditional probability of seeing noun n_1 as a modifier of noun n_2 augmented with a dispersion parameter for n_1 (see equation (6.11)).

$$IC(n_1,[n_1,n_2]) = \frac{f(n_1,n_2)}{n_{n_1} + d_{n_1} - 1} \tag{6.11}$$

Here, $f(n_1,n_2)$ is the frequency of the sequence n_1,n_2 in the corpus, n_{n_1} is the number of noun-noun pairs in which n_1 occurs in the first position, and d_{n_1} is a dispersion parameter

[1] Note that informational contribution is defined in terms of conditional probability and has no information-theoretic interpretation (Shannon and Weaver 1949) as its name misleadingly suggests.

Table 6.3: Example values for co-occurrence frequency, conditional probability, informational contribution, mutual information and the log-likelihood ratio

	world war	*term shilling*
CoocF	3,707	1
CondP	.12455	.00098
IC	.29663	.00036
MI	9.00302	1.63
LLRatio	32,301.15	.91

which can be understood as the number of distinct head nouns with which n_1 is paired. When $IC(n_1,[n_1,n_2]) = 0$, n_1 and n_2 never co-occur, whereas when $IC(n_1,[n_1,n_2]) = 1$, n_1 co-occurs only with n_2. As shown in Table 6.3, the informational contribution value (IC) for *world war* is approximately 10^3 times greater than the IC value for *term shilling*.

Mutual Information. Mutual information has been widely used to discover word associations (e.g., Church and Hanks 1990) as well as collocations (e.g., Lin 1999). We employ mutual information as a measure of the likelihood of observing nouns n_1 and n_2 together more often than expected by their individual frequencies.[2]

$$(6.12) \quad I(n_1,n_2) = \log_2 \frac{P(n_1,n_2)}{P(n_1)P(n_2)}$$

Here, $P(n_1,n_2)$ is estimated by the number of times n_1 and n_2 co-occur divided by the size of the corpus N and similarly $P(n_1)$ and $P(n_2)$ are estimated by the number of times n_1 and n_2 are observed in the corpus divided again by N. The amount of information we gain about *war* increases by 9 bits if we know that it is preceded by *world*. On the other hand, the amount of information we have about *shilling* when preceded by *term* is only increased by 1.6 bits (see Table 6.3). This means that the likelihood of observing *world* with *war* is greater than the likelihood of observing *term* and *shilling* together.

Log-likelihood ratio. It has been previously shown that the log-likelihood ratio (G-score) of an expression is a good indicator of its terminological status (Daille 1996). Furthermore, as demonstrated by Dunning (1993) the statistic adequately takes into account the frequency of the co-occurring words and is less sensitive than mutual information or the χ^2 statistic to rare events and corpus size (see Dunning 1993 for details on how to calculate the log-likelihood ratio). We expect noun-noun sequences which are valid compounds such as *world war* to have high log-likelihood values and non-compounds such as *term shilling* to have low log-likelihood

[2]Note that we are making use of *pointwise mutual information*, originally defined by Fano (1961: 27), which in our case measures how much information we have about word n_2 when we see word n_1 in the corpus or vice versa. Fano's (1961) definition of mutual information differs from the standard notion of mutual information which is defined between two random variables X and Y and not between two particular events x and y.

values (see Table 6.3).

We investigated whether any of the above statistical scores can discover compounds and reliably distinguish them from non-compounds by running three experiments which examined the behavior of the five corpus-based measures under varying assumptions with respect to the frequency of the candidate compounds. In Experiment 11 (see Section 6.4) we test the behavior of the five scores on a sample containing hapaxes (CoocF ≥ 1). Experiment 12 (see Section 6.5) replicates Experiment 11 on a sample which does not contain hapaxes (CoocF > 1). In Experiment 13 (see Section 6.6) we observe the performance of the five scores on a sample which contains only frequent noun-noun sequences (CoocF ≥ 5).

6.4. Experiment 11: Evaluation of Statistical Scores (CoocF ≥ 1)

6.4.1. Method

We computed the mutual information (MI), conditional probability (CondP), informational contribution (IC), co-occurrence frequency (CoocF), and the log-likelihood ratio (LLRatio) for all noun-noun sequences contained in the sample on which our extraction procedure attained a precision of 71.0% (870 tokens, see Section 6.2.1). One would expect an ideal score to assign higher values to the 71.0% of the noun-noun sequences which are valid compounds and lower values to the erroneous 29.0%.

We also measured recall as an indicator of the type of compounds (i.e., established versus novel) contained in the sample. We compared the compounds in our sample against a dictionary of compound nouns (81,246 entries) which was compiled from four sources: WordNet (Miller et al. 1990), CELEX (Burnage 1990), a list of compound nouns from a spell-checker for the NEXT computer compiled by George Fowler, and finally a list of compound nouns collected by Richard Sproat (Sproat 1994). Recall was measured as the number of compounds found both in the sample and compound noun dictionary over the number of compounds contained in the sample. In the experiments described below we use recall as an indicator of compound noun productivity rather than a measure of the number of compounds found by any given statistic out of the number of compounds there are. Compounding is a very productive phenomenon and a large number of compounds, especially novel ones, are not to be found in pre-existing dictionaries which typically focus on established compounds. In other words, we expect recall to be relatively low due to the productivity of compound noun formation and also because dictionaries are not perfect sources of compounds, even in cases where these are established ones.

6.4.2. Results

The histograms in Figure 6.1 plot the values of the five corpus-based measures (x axis) against their accuracy in detecting noun-noun compounds (y axis). More specifically, the range of values for all five statistics was split into 70 equal-sized intervals (bins). The values for each statistical score were plotted against precision (i.e., number of valid compounds out of candidate compounds in the bin, y axis).

As shown in Figure 6.1 the precision for mutual information (see Figure 6.1a), conditional probability (see Figure 6.1b) and informational contribution (see Figure 6.1c) fluctuates. This means that we cannot establish any relationship between the values of these three statistics and their precision. Also note that the precision for noun-noun sequences falling in the highest bin is low for mutual information (28.57%) and conditional probability (18.18%). The precision histograms for co-occurrence frequency (see Figure 6.1d) and the log-likelihood ratio (see Figure 6.1d) display less fluctuations and thus enable us to determine a threshold above which we can reliably acquire compound nouns (e.g., CoocF $\geq$ 5 and the LLRatio $\geq$ 45). These results are compatible with previous work in terminology acquisition (see Daille 1996 and Manning and Schütze 1999 for overviews), where it has been shown that co-occurrence frequency and log-likelihood ratio are better indicators of the termhood of a given word combination than statistics such as mutual information or conditional probability. A well known property of MI is its tendency to overemphasize rare events (Church and Hanks 1990): 83.3% of the candidate compounds which are assigned the highest MI values (MI $>$ 19) are attested in the corpus only once. This seems to be also true for IC and CondP: 91.2% of the candidate compounds assigned the highest IC value (IC $=$ 1) and 81.8% of the candidate compounds with the highest CondP value (CondP $=$ 1) are attested in the corpus once.

However, note that despite the fact that CoocF and LLRatio are better suited for discovering compounds than IC, MI, or CondP, they do not fully succeed in distinguishing between valid compounds and nonce noun sequences. Although noun-noun sequences with high CoocF and LLRatio values are unambiguously compounds, noun-noun sequences with small frequencies do not strictly correspond to nonce terms. Note that 66.0% of the noun-noun sequences falling in the lowest CoocF values and 68.6% of the sequences with the lowest LLRatio values are valid compounds (see Figures 6.1d and 6.1e).

Only 5.8% of the compounds contained in our sample were found in the compound noun dictionary (see Table 6.4). This means that a substantial number of compounds are not established, but novel. This is not surprising because compounding is highly productive. As mentioned in Section 6.1, compounds can be created to serve as *deictic markers* in order to satisfy the speaker's communicative purposes (Downing 1977).

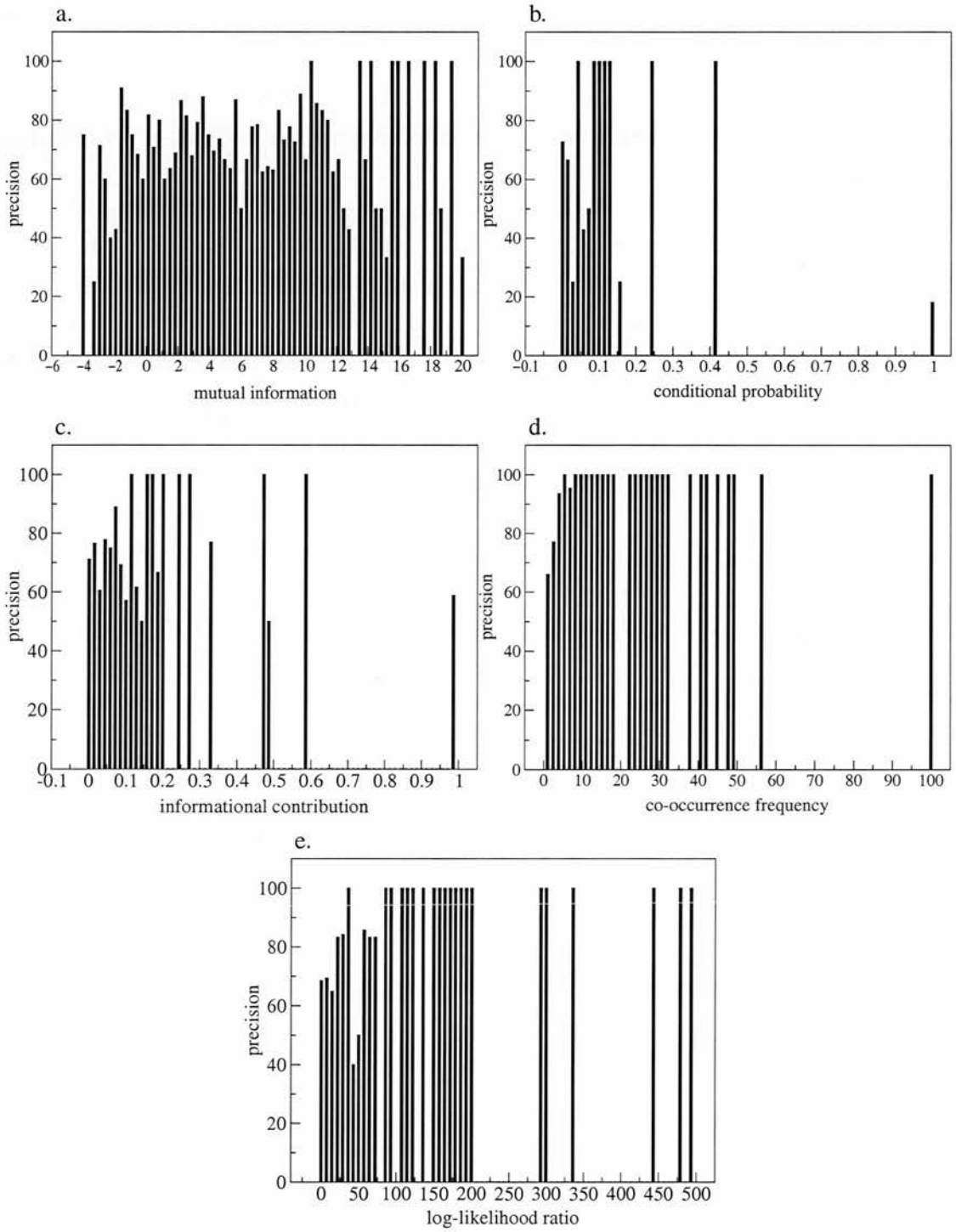

Figure 6.1: Precision of mutual information, conditional probability, informational contribution, co-occurrence frequency, and the log-likelihood ratio on manually annotated data (CoocF ≥ 1)

Table 6.4: Relation of compound noun co-occurrence frequency with precision and recall

CoocF	BNC	Sample	Precision (%)	Recall (%)
≥ 5	52,832	538	93.5	19.1
> 1	160,214	545	82.0	9.8
≥ 1	510,673	871	71.0	5.8
$= 1$	350,459	532	57.7	1.0

6.5. Experiment 12: Evaluation of Statistical Scores (CoocF > 1)

6.5.1. Method

We further examined the behavior of the five scores by discarding from the data all noun-noun sequences with co-occurrence frequency of one. This reduced the number of potential compounds by a factor of three (see Table 6.4). A random sample (545 tokens) was selected from this population. The noun-noun sequences in the sample were manually classified as compounds or not within context using Xkwic (Christ 1995). 82.0% of the sequences contained in the sample were valid compounds (see Table 6.4). This means that discarding hapaxes alone results in a precision increase of 11.0%. We computed the five corpus-based scores for all noun-noun sequences in the sample and plotted their values against their precision as described above (see Section 6.4.1). Recall was measured as described in Section 6.4.1.

6.5.2. Results

Mutual information, conditional probability and informational contribution show patterns similar to the histograms in Figure 6.1 (see Figures 6.2a–6.2c). Notice that in this sample co-occurrence frequency does slightly worse than log-likelihood ratio (compare Figures 6.2d and 6.2e). Even though hapaxes (i.e., candidate compounds with CoocF $= 1$) are absent, neither co-occurrence frequency nor the log-likelihood ratio can reliably distinguish false positives (i.e., non-compounds). Noun-noun sequences with a low LLRatio or CoocF value are likely to be compounds. Discarding hapaxes yields an increase in precision even for noun-noun sequences with very small frequencies: 75.8% of the candidate noun-noun sequences with co-occurrence frequency of two are valid compounds. This, however, considerably limits the number of potential compounds to be identified since two thirds of the candidate noun-noun sequences are ignored (see Table 6.4).

As shown in Table 6.4 an increase of 4.0% in recall is achieved when hapaxes are eliminated. This increase is relatively small considering that only 31.4% of the candidate compounds extracted from the corpus are taken into account. This means that a small number of established compounds is retrieved.

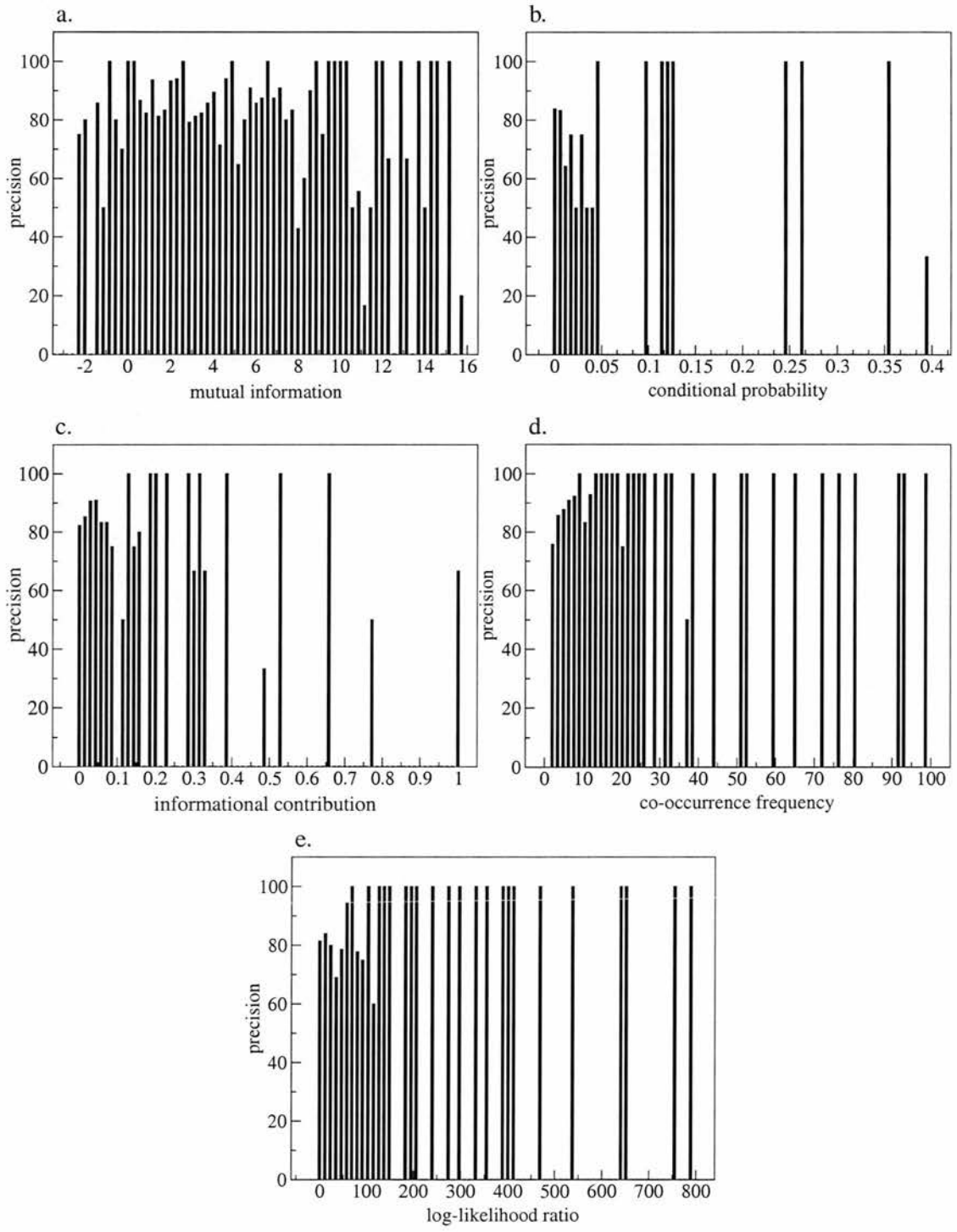

Figure 6.2: Precision of mutual information, conditional probability, informational contribution, co-occurrence frequency, and the log-likelihood ratio on manually annotated data (CoocF > 1)

6.6. Experiment 13: Evaluation of Statistical Scores (CoocF $\geq$ 5)

6.6.1. Method

An interesting question is whether the behavior of the corpus-based statistics changes when the co-occurrence frequency of the noun-noun sequence is not very small. In order to investigate this we discarded from the data all noun-noun sequences with a co-occurrence frequency smaller than five. The number of candidate compounds was drastically reduced by a factor of ten (see Table 6.4). From the remaining sequences we selected and manually disambiguated (i.e., classified a noun-noun sequence as a valid compound or not) a random sample of 538 tokens, using Xkwic (Christ 1995). 93.5% of the noun-noun sequences in this sample were valid compounds. We computed the five corpus-based scores for all noun-noun sequences in the sample and plotted their values against their precision as described in Section 6.4.1. Recall was measured as described in Section 6.4.1.

6.6.2. Results

Figure 6.3 shows how precision varies with respect to different values for the five corpus-based measures. Note that the precision for mutual information and informational contribution fluctuates even with a co-occurrence frequency larger than four (see Figures 6.3a,c). This confirms that these two scores are inappropriate for compound-noun acquisition. In contrast to Figures 6.1b and 6.2b, the precision for conditional probability shows considerably less fluctuation (see Figure 6.3b) and thus enables us to reliably acquire compounds above a certain threshold (e.g., CondP $>$ 0.04). The behavior of both co-occurrence frequency and the log-likelihood ratio remains stable.

In sum, we conclude that both co-occurrence frequency and log-likelihood ratio are good indicators of valid compounds. Conditional probability can be also used, although only when the co-occurrence frequency of the noun-noun sequence is larger than four. A general observation is that all five statistics, the log-likelihood ratio and co-occurrence frequency included, fail to indicate a clear-cut distinction between compounds and non-compounds.

Recall attains its highest value (19.1%) when CoocF $\geq$ 5 (see Table 6.4). This suggests that a greater number of established compounds will be acquired with larger co-occurrence frequency values.

6.7. Experiment 14: Hapaxes

6.7.1. Method

We further examined the properties of hapaxes and in particular how likely are noun-noun sequences with co-occurrence frequency of one to form valid compounds. We concentrated solely

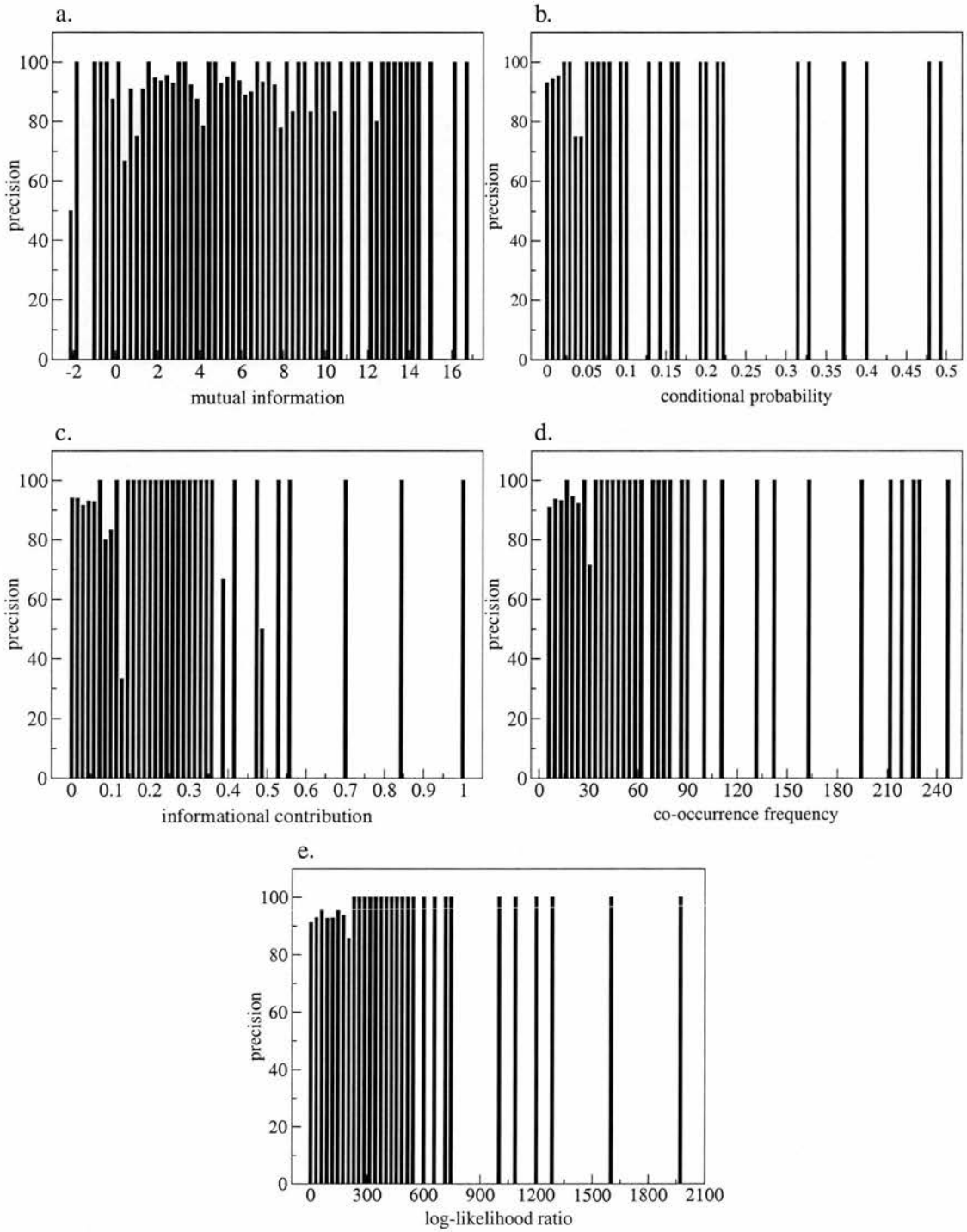

Figure 6.3: Precision of mutual information, conditional probability, informational contribution, co-occurrence frequency, and the log-likelihood ratio on manually annotated data (CoocF $\geq$ 5)

on hapaxes by eliminating from the noun-noun sequences retrieved from the BNC all candidate compounds with a co-occurrence frequency greater than one. This resulted in 350,459 potential compounds from which we selected a random sample of 532 tokens (see Table 6.4). The noun-noun sequences in the sample were manually disambiguated using Xkwic (Christ 1995). Precision was measured as the number of valid compounds contained in the sample. Recall was measured as described in Section 6.4.1.

6.7.2. Results

57.7% of the sequences in this sample were valid compounds (see Table 6.4). Note further that hapaxes represent 68.6% of the candidate noun-noun sequences retrieved from the BNC (350,459 out of 510,673). None of the statistical scores introduced in Section 6.3 can be used to distinguish rare valid compounds from rare nonce terms. Co-occurrence frequency relies on the assumption that repetition is a good indicator for distinguishing terms from non-terms (Justeson and Katz 1995c). The log-likelihood ratio applies hypothesis testing to term discovery on the basis of the assumption that two words form a term when they co-occur more often than chance (Daille 1996; Dunning 1993). In the case of hapaxes, the distributional properties of compounds do not differ considerably from the distributional properties of non-compounds and consequently, not even the log-likelihood ratio, a score which otherwise seems well suited for discovering compounds, can be used to distinguish them. The noun-noun sequences given in Table 6.5 are all examples of a valid compound and a non-compound which have identical log-likelihood values.

We mentioned earlier that 57.7% of the noun-noun sequences occurring in the corpus only once are valid compounds. Of the remaining non-valid compounds (42.3%), 61.9% are tagging errors (i.e., if tagging was perfect these sequences would have been excluded), 30.6% are due to the absence of structural information (i.e., they would have been ruled out if parsing information was available), 5.30% are acronyms, and 2.20% are foreign terms or typographical mistakes. Let us concentrate on the tagging mistakes which are the most frequent misclassifications. The BNC was tagged automatically, using CLAWS4, a probabilistic part-of-speech tagger, with error rate ranging from 3% to 4% (Leech et al. 1994). A particular feature of this tagger is that it assigns ambiguity tags (also called "portmanteau tags") in cases where it is unable to assign a category with certainty. Although 46.3% of the tagging misclassifications are noun sequences with ambiguity tags, discarding candidate compounds with ambiguity tags would ignore a substantial number of valid compounds (35.6%), since ambiguity tags are a feature of both valid compounds and nonce noun-noun sequences. Consider the sentences (6.13a) and (6.14a): both sequences *painting tours* and *proof falls* bear the ambiguity tag NN2-VVZ (NN2 stands for plural nouns and VVZ for third person singular verbs); the same holds for the candidate sequences *formulation meeting* and *mist dispersing* in (6.14) which have the ambiguity tag NN1-VVG (NN1 stands for singular nouns and VVG for the *-ing* form of verbs).

Table 6.5: Compounds and non-compounds with identical log-likelihood values

LLRatio	Compound	Non-compound
0.02	librarian problem	city hundred
0.34	form rating	chance second
2.59	stove firebreak	effect chair
3.83	network documentary	rate effort
4.11	incompetence argument	submarine green

(6.13) a. He had previously made visits to Egypt and the Holy Land and undertaken several walking and [painting$_{\text{NN1}}$ tours$_{\text{NN2-VVZ}}$] in the Alps with his wife.

b. Perhaps she will attend a [formulation$_{\text{NN1}}$ meeting$_{\text{NN1-VVG}}$] by the time she's two.

(6.14) a. The burden of [proof$_{\text{NN1}}$ falls$_{\text{NN2-VVZ}}$] on the party seeking to enforce the restraint.

b. The weather was clearing, the [mist$_{\text{NN1}}$ dispersing$_{\text{NN1-VVG}}$], the sun began to peep out.

Recall from Section 6.2.2 that Lauer's (1995) study concentrated on nouns that did not exhibit part-of-speech ambiguity: the corpus used for the extraction of compounds was not annotated with parts of speech, instead noun-noun sequences were identified from a predefined set of unambiguous nouns. We showed that a considerably lower accuracy (71.0% versus Lauer's 97.9%) is attained when candidate compounds are extracted from domain-independent unrestricted text, annotated with part-of-speech information. We have found that 68.6% of the candidate compounds extracted from the BNC are hapaxes, 57.7% of which are valid compounds.

Finally, recall reaches its lowest value for hapaxes (1.0%, see Table 6.4). A general observation is that recall increases with co-occurrence frequency. Assuming that recall is an indicator of the proportion of novel compounds in the data (i.e., the higher the recall the more lexicalized compounds found in the sample), we can conclude that the largest number of novel compounds is to be expected with small co-occurrence frequency values. Novel compounds are typically speaker and situation dependent, and therefore tend to be low in frequency.

6.8. Discussion

We showed that a simple heuristic which looks for consecutive nouns in the BNC (see Section 6.2.2) results in substantially lower accuracy (i.e., 71.0%) than previously reported in the literature (i.e., 97.9%, Lauer 1995). This discrepancy indicates that existing methodologies are domain and corpus specific and do not easily scale up to large diverse corpora such as the BNC. Co-occurrence frequency and the log-likelihood ratio allow us to establish thresholds above which we can consider candidate noun-noun sequences as compounds with varying degrees of accuracy (see Table 6.4). Conditional probability can be used only when the co-occurrence

frequency of the candidate compound is larger than four. Other corpus-based measures such as mutual information and informational contribution are inappropriate for the acquisition of compound nouns even when the frequency of the noun-noun pair is high. High thresholds yield compounds with a high degree of accuracy (e.g., 93.5% of noun-noun sequences with co-occurrence frequency larger than four are valid compounds). This means, however, that only a small fraction of the compounds attested in the corpus is acquired.

Our evaluation revealed that hapaxes are the majority of the candidate compounds extracted from the corpus (see Section 6.7). 57.7% of these noun-noun sequences are valid compounds, a large number of which are novel (i.e., created on the spot to satisfy the speaker's communicative needs) and cannot be distinguished from nonce terms solely on the basis of their distributional properties.

In the next sections we turn to hapaxes and propose a method that distinguishes valid compounds from nonce noun sequences by modeling the distributional tendencies observed in lexicalized (i.e., frequent) compounds. We use corpus evidence about established compounds to estimate features that can discriminate rare valid compounds from rare nonce terms. We introduce several linguistic features as potential indicators of the compounding relation between two nouns such as the likelihood of a given word as a compound head or modifier, the context surrounding the candidate compound, and the likelihood of two nouns to form a meaningful compound. We explore their individual and combined contribution using decision tree learning. In Section 6.9.1 we present and motivate these features. Section 6.9.4 details our machine learning experiments and Section 6.10 discusses our results.

6.9. Experiment 15: Decision Tree Learning

6.9.1. Features for Discovering Compounds

In what follows we present a method that attempts to distinguish compounds from non-compounds in cases where the assumptions underlying lexical-association scores do not hold. In particular, we show how an approach which uses decision trees and linguistically motivated features can achieve satisfactory performance at distinguishing compounds from non-compounds in the case of noun-noun sequences attested in the corpus only once. Recall from the results of Experiment 14 (see Section 6.7) that hapaxes constitute the majority (68.6%) of all noun-noun sequences attested in the corpus and furthermore that more than half of these are valid compounds (see Table 6.4).

We combine evidence from different information sources in order to classify a given noun-noun sequence as a valid compound or a nonce noun-noun sequence. In our machine learning experiments we make use of numerical features (i.e., frequency, probability) as well as categorial features (i.e., the context of the candidate noun-noun sequence) and explore their contribution in the classification task. We estimate our numerical features by making use of

Table 6.6: Feature values for noun-noun sequences

	$f(n_1)$	$f(n_2)$	$P(\mathrm{H}\|n_1)$	$P(\mathrm{M}\|n_2)$	$f(c_1,c_2)$
cocaine customer	71	159	1	.18	285.85
baby calf	740	22	.91	.15	35.13
people excitement	1,823	9	.45	1	4.98
may push	0	35	0	.43	76.93

previously attested compounds. More specifically, we estimate the numeric quantities detailed below from a corpus consisting of all noun-noun sequences with BNC co-occurrence frequency greater than four (see Table 6.4). Note that 93.5% of these sequences are valid compounds and can therefore provide useful information about the likelihood of a given noun as a compound head or modifier. In the following we describe and motivate these features:

Noun frequency. Given a noun-noun sequence n_1n_2 we look at whether the frequency of the head n_2, $f(n_2)$, or the frequency of the modifier n_1, $f(n_1)$, are reliable indicators for distinguishing compounds from non-compounds. As mentioned earlier we estimate the frequencies $f(n_1)$ and $f(n_2)$ from a corpus of noun-noun sequences (52,832 in total) that are attested in the BNC with co-occurrence frequency greater than four. Consider for example the compound *cocaine customer* which is attested in the BNC only once (see Table 6.6). The word *cocaine* is attested as a modifier 71 times and the word *customer* is attested as a head 159. Compare now *cocaine customer* to *people excitement* which is not a valid compound and is also found in the BNC once (the sequence is attested in the sentence *For some people excitement is only possible outside marriage.*). The modifier frequency $f(people)$ is 1,823 whereas the head frequency $f(excitement)$ is nine. Clearly, *excitement* is less likely to be a compound head when compared to *customer* (see Table 6.6).

Probability. Given a noun-noun sequence n_1n_2 we investigate whether it is likely for n_2 to be a head and for n_1 to be a modifier. We express these quantities as follows:

$$P(\mathrm{M}|n_2) = \frac{f(\mathrm{M},n_2)}{f(n_2)} \tag{6.15}$$

$$P(\mathrm{H}|n_1) = \frac{f(n_1,\mathrm{H})}{f(n_1)} \tag{6.16}$$

Here, $f(\mathrm{M},n_2) = \sum_{n_1} f(n_1,n_2)$ and $f(n_1,\mathrm{H}) = \sum_{n_2} f(n_1,n_2)$. Equation (6.15) expresses the likelihood of n_2 as a head (preceded by a noun modifier) and equation (6.16) expresses the likelihood of n_1 as a modifier (followed by a noun head). We estimate $f(\mathrm{M},n_2)$ and $f(n_1,\mathrm{H})$ from the reliable noun-noun sequences attested previously in the corpus (CoocF ≥ 5). The frequencies $f(n_1)$ and $f(n_2)$ are the number of times we see n_1 and n_2 in our estimation corpus

independently of their position (i.e., independently of whether they are heads or modifiers).

Consider the compounds *cocaine customer* and *baby calf* in Table 6.6. The likelihood of the words *cocaine* and *baby* to be found in a modifier position is very high (1 and .91, respectively). Contrast this with the sequence *may push* which is the result of a tagging mistake (i.e., both *may* and *push* are annotated as nouns in the sentence *Their different responsibilities in relation to the public may push them in opposite directions*): the likelihood of the word *may* to be found in a modifier position is zero. Note further that *push* can be a noun (denoting the act of pushing) and therefore it is not entirely unlikely to be found in a head position (see Table 6.6).

Concept frequency. The features described above do not capture meaning regularities concerning the compounding process. It has been argued that generative devices in the lexicon are responsible for inducing predictable sense alternations (Pustejovsky 1995). This means that there are classes of compounds generated from rules capturing their systematic properties (Copestake and Lascarides 1997; Jones 1995; Warren 1978). For example the compounds *metal tube*, *leather belt*, and *tin cup* are the result of a rule that combines a substance with an artefact to produce a compound noun that denotes an artefact *made of* a substance. The rule which combines a location with an entity (either animate or inanimate) produces the compounds *country boy* and *hospital bill*. Such rules encode linguistic constraints predicting thus why *belt leather* and *boy country* are odd under the interpretations "leather made of belt" and "country from boy".

For the purposes of our experiments it would be useful to have some notion of how likely a certain concept combination is. Consider again the sequences *cocaine customer* and *people excitement* from Table 6.6. We would expect the combination of the concepts representing *cocaine* and *customer* to be more frequent than the combination of the concepts representing *people* and *excitement*. A straightforward way to capture this would be by substituting the head and modifier by the concepts with which they are represented in a taxonomy (i.e., a hierarchy of concepts organized in terms of hypernymic and hyponymic relations). So the frequency of the concept combination $f(c_1, c_2)$ could be estimated by counting the number of times the concept c_1 corresponding to n_1 was observed as the modifier of the concept c_2 corresponding to the head n_2.

This would be a straightforward task if each word was always represented in the taxonomy by a single concept or if we had a corpus of compounds labeled explicitly with taxonomic information. Lacking such a corpus we need to take into consideration the fact that words in a taxonomy may belong to more than one conceptual class. Nouns in WordNet (Miller et al. 1990) correspond to an average of 11.5 concepts (e.g., the word *return* belongs to 104 distinct conceptual classes), whereas nouns in Roget's thesaurus correspond to an average of 1.7 concepts (e.g., the word *point* has 18 distinct concepts). Because a head or a modifier can generally be the realization of one of several conceptual classes, counts of modifier-head configurations

Table 6.7: Estimated concept pair frequencies

$\langle c_1, c_2 \rangle$	$f(c_1, c_2)$	Examples
⟨substance, object⟩	604.7	iron table
⟨act, social group⟩	403.0	mining family
⟨entity, location⟩	382.4	girls school
⟨group, relation⟩	267.6	world language
⟨communication, act⟩	231.1	speech treatment
⟨person, artefact⟩	162.1	developer's kit
⟨time period, psychological feature⟩	75.6	autumn joy
⟨institution, person⟩	38.7	bank spokesman

must be constructed for all potential concept combinations.

To give a concrete example consider again the compound *cocaine customer*. The word *cocaine* has one sense in WordNet and belongs to six conceptual classes (⟨hard drug⟩, ⟨narcotic⟩, ⟨drug⟩, ⟨artefact⟩, ⟨object⟩, ⟨entity⟩). The word *customer* has also one sense in WordNet and belongs to five conceptual classes (⟨consumer⟩, ⟨person⟩, ⟨life form⟩, ⟨causal agent⟩, ⟨entity⟩). Since we do not know which particular instantiation of these conceptual classes *cocaine* and *customer* are, we will distribute the attested frequency of *cocaine customer* over all pairwise concept combinations. We formally define the set of concept combinations as follows:

$$c(n_1, n_2) = \{\langle c_i, c_j \rangle \mid c_i \in classes(n_1), c_j \in classes(n_2), c_i \neq c_j\} \tag{6.17}$$

Here, $c(n_1, n_2)$ is the set of distinct concept pairs a given noun-noun sequence is an instantiation of. Note that we impose a restriction on the type of concept pairs we generate, namely we disallow pairs with identical concepts (see (6.17)). The motivation for this restriction is twofold: first, we want to avoid overly general concept pairs that could potentially represent any noun-noun combination (e.g., ⟨entity, entity⟩, ⟨artefact, artefact⟩); second, it is implicitly assumed in the theoretical linguistics literature (Levi 1978; Warren 1978) that compounds are derived through combinations of distinct concepts. The pairs $\langle c_i, c_j \rangle$ can be thought of as rules imposing constraints on concept combinations.

For each compound in our corpus we generate the set of concept pairs it is potentially an instantiation of. The compound *cocaine customer* generates 29 concept pairs (e.g., ⟨artefact, consumer⟩, ⟨artefact, person⟩, etc.). We estimate the frequency of a concept pair $f(c_1, c_2)$ by summing over all noun-noun sequences $n_1 n_2$ that are representative of the concept combination $\langle c_1, c_2 \rangle$. We divide the contribution of each compound $n_1 n_2$ by the

number of concept combinations it represents (Lauer 1995; Resnik 1993):

$$f(c_1,c_2) \approx \sum_{\langle n_1,n_2\rangle \in \langle c_1,c_2\rangle} \frac{f(n_1,n_2)}{|c(n_1,n_2)|} \quad (6.18)$$

Here, $f(n_1,n_2)$ is the number of times a given compound was observed in the corpus (i.e., noun-noun sequences with CoocF ≥ 5) and $|c(n_1,n_2)|$ is the number of conceptual pairs n_1n_2 has. Assuming that we want to take the compound *cocaine customer* into account for estimating the frequency of the concept pair $\langle$`artefact,person`$\rangle$, we will increment the observed co-occurrence count of $\langle$`artefact,person`$\rangle$ by $\frac{1}{29}$, since *cocaine customer* is represented by 29 distinct concept pairs. Table 6.7 shows a random sample of the derived concept pairs and their estimated frequencies.

Assume now that we want to decide whether the sequence *people excitement* is a valid compound or not. We generate all pairs of conceptual classes represented by *people excitement* (see (6.17)). The word *people* has four senses and belongs to 6 conceptual classes; *excitement* has also four senses and belongs to 15 classes. This means that *people excitement* is potentially represented by 90 concept pairs (*people* and *excitement* have no concepts in common), the frequency of which can be estimated from our corpus of valid compounds using (6.18). Since we do not know which are the actual classes for the nouns *people* and *excitement* in the corpus, we weight the contribution of each class pair by taking the average of the estimated frequencies for all 90 class pairs:

$$f(n'_1,n'_2) = \frac{\sum_{\langle c_1,c_2\rangle \in c(n'_1,n'_2)} f(c_1,c_2)}{|c(n'_1,n'_2)|} \quad (6.19)$$

As shown in Table 6.6 *people excitement* is much less likely than *cocaine customer*. Also note that *may push* is considered fairly likely (in fact more likely than *baby calf* which is a valid compound) since both *May* and *push* can be nouns and are listed as such in the WordNet taxonomy. Note that the estimation of the concept frequencies in (6.18) relies on the simplifying assumption that a given noun is equally likely to be represented by any of its conceptual classes. As a result, the occurrence frequency of a compound is evenly distributed across all possible concept combinations representing the nouns forming the compound, since we cannot assess (without access to a corpus annotated with class information) which concept combinations are likely and which are not.

Context. Although the numerical features described above encode important information with respect to modifier-head relations and their properties, they are blind to contextual or part-of-speech information that can help detect parsing or tagging errors. Recall from Section 6.7 that 42.3% of the noun-noun sequences that occur in the corpus only once are not valid compounds. Of these 96.1% are due to tagging errors and to the absence of structural information

(i.e., parsing). Contextual information could help detect noun-noun sequences that do not stand in a modification relation but happen to be adjacent or correct for erroneous classifications in the case of tagging mistakes.

Consider again the noun-noun sequence *may push* from Table 6.6, which is attested in sentence (6.20a). In this case, the context strongly indicates that *may push* is not a compound given that *push* is followed by a pronoun. Pronouns (e.g., personal pronouns) typically precede compound nouns but never follow them. We encode contextual information as the number of words preceding and succeeding the noun-noun sequence in question. In order to capture grammatical and syntactic dependencies we reduce words to their parts of speech and encode their specific positions to the left or right of the candidate compound. An example of this type of feature-encoding is given in (6.20b) which represents the context surrounding *may push* in sentence (6.20a). The feature-vector in (6.20b) consists of the candidate compound *may push*, represented by its parts of speech (NN1 and NN1, respectively) and a context of four words to its right and four words to its left, also reduced to their parts of speech.[3]

(6.20) a. Their different responsibilities in relation to the public may push them in opposite directions.

b. [NN2, PRP, AT0, AJ0, NN1, NN1, PNP, PRP, AJ0, NN2]

In the following we explore how the two types of features (i.e., numerical and categorial) perform independently as well as in combination. We combine these distinct information sources using Ripper, a decision tree classifier (Cohen 1996). Ripper induces classification rules from a set of preclassified examples. In our experiments Ripper was trained on 631 noun-noun sequences randomly selected from the BNC and tested on 200 unseen sequences. All candidate compounds in the training and test data were hapaxes.

6.9.2. Agreement

The data reserved for testing the performance of the rule learner was also used to evaluate whether humans can reliably distinguish compounds from non-compounds. The task assessed both the quality of the test data and the difficulty of the task. If humans do not agree in their classifications of candidate compounds, there is little hope for our machine learning methods.

Two judges decided whether a candidate noun-noun sequence was a compound. The judges were given a page of guidelines but no prior training (the guidelines are given in Appendix A, Section A.2). The candidate compounds were classified in context: the judges were given the corpus sentence in which the noun-noun sequence occurred together with the previous and following sentence. We measured the judges' agreement on the classification task using the Kappa coefficient (Cohen 1960, see Section 2.5.1).

[3]The part-of-speech NN1 stands for singular common nouns, NN2 stands for plural common nouns, AT0 stands for determiners, PRP for prepositions, PNP for pronouns, and AJ0 for adjectives.

The judges' agreement on the disambiguation task was $K = .80$ ($N = 200$, $k = 2$). The agreement was good given that the judges had minimal instructions and no prior training. In the following section we report the results of our machine learning experiments on classifying candidate compounds.

6.9.3. Method

We explored the contribution of the features detailed in Section 6.9.1 in classifying candidate compounds as valid or nonce terms. 54.5% of the noun-noun sequences in the test data were valid compounds. This means that a heuristic which simply defaults to classifying any noun-noun sequence as a compound achieves a precision of 54.5%. We investigated how the different types of features influence the classification task. In particular, we examined: (a) whether simple numeric features such as frequency and probability can achieve a better performance than the default strategy, (b) whether taxonomic information increases classification accuracy, (c) whether contextual information alone is a good indicator for valid compounds, and (d) whether numerical and categorial features are complementary.

As mentioned in Section 6.9.1 the numerical features were estimated from the corpus of valid compounds acquired from the BNC (CoocF ≥ 5). For the estimation of the concept frequency feature (see Section 6.9.1), we experimented with two concept hierarchies, Roget's thesaurus and WordNet (see Section 2.2 for details). In what follows the learner's output is compared against the manual classification (see Section 6.9.2) and accuracy is measured accordingly.

6.9.4. Results

In Table 6.8 we show how classification performance varies using individual numerical features. For comparison we also show the performance of the simple strategy of always classifying a noun-noun sequence as a compound (D). The best feature is concept frequency using WordNet with an accuracy of 72.0%. We used the χ^2 statistic to examine whether the observed performance was significantly better than the simple strategy of always defaulting to a valid compound classification which yields an accuracy of 54.5%. The proportion of noun-noun sequences classified correctly was significantly greater than 54.5% for the concept frequency feature using WordNet and Roget ($p < .01$ and $p < .05$, respectively) and for $P(\mathrm{H}|n_1)$, the feature encoding the likelihood of a noun as a modifier ($p < .05$).

Table 6.9 reports results on the classification task using pairs of the six numeric features. When no taxonomic information is used, the best performance (68.0%) is achieved by combining the features $f(n_2)$, the frequency of the head noun, and $P(\mathrm{H}|n_1)$, the probability of encountering a head given that we have seen a noun in the modifier position. The proportion of compounds classified correctly using these two features was significantly greater than 54.5%

Table 6.8: Individual numeric features

Features	Train (%)	Test (%)
D	56.9 ± 1.97	54.5 ± 3.53
$f(n_1)$	61.5 ± 1.94	60.5 ± 3.47
$f(n_2)$	57.4 ± 1.97	59.0 ± 3.49
$P(\mathrm{H}\|n_1)$	61.2 ± 1.94	64.0 ± 3.40
$P(\mathrm{M}\|n_2)$	61.9 ± 1.94	61.0 ± 3.46
$f_{wn}(n_1,n_2)$	66.4 ± 1.88	72.0 ± 3.18
$f_{ro}(n_1,n_2)$	60.1 ± 1.94	66.5 ± 3.35

Table 6.9: Decision trees with two numerical features

Features	Train (%)	Test (%)
D	56.9 ± 1.97	54.5 ± 3.53
$P(\mathrm{H}\|n_1)$, $P(\mathrm{M}\|n_2)$	63.4 ± 1.95	62.0 ± 3.44
$P(\mathrm{H}\|n_1)$, $f(n_1)$	61.2 ± 1.94	64.0 ± 3.40
$P(\mathrm{H}\|n_1)$, $f(n_2)$	65.8 ± 1.93	68.0 ± 3.31
$P(\mathrm{M}\|n_2)$, $f(n_1)$	62.3 ± 1.93	61.0 ± 3.46
$P(\mathrm{M}\|n_2)$, $f(n_2)$	61.5 ± 1.94	61.0 ± 3.46
$f(n_1)$, $f(n_2)$	60.1 ± 1.95	62.0 ± 3.44
$f_{wn}(n_1,n_2)$, $P(\mathrm{H}\|n_1)$	66.4 ± 1.88	72.0 ± 3.18
$f_{wn}(n_1,n_2)$, $P(\mathrm{M}\|n_2)$	66.4 ± 1.88	72.0 ± 3.18
$f_{wn}(n_1,n_2)$, $f(n_1)$	66.4 ± 1.88	72.0 ± 3.18
$f_{wn}(n_1,n_2)$, $f(n_2)$	66.2 ± 1.88	72.5 ± 3.17
$f_{ro}(n_1,n_2)$, $P(\mathrm{H}\|n_1)$	62.4 ± 1.93	67.0 ± 3.33
$f_{ro}(n_1,n_2)$, $P(\mathrm{M}\|n_2)$	62.6 ± 1.93	67.5 ± 3.32
$f_{ro}(n_1,n_2)$, $f(n_1)$	61.3 ± 1.94	66.5 ± 3.35
$f_{ro}(n_1,n_2)$, $f(n_2)$	64.5 ± 1.91	66.5 ± 3.35
$f_{wn}(n_1,n_2)$, $f_{ro}(c_1,c_2)$	66.4 ± 1.88	72.0 ± 3.18

($p < .01$). This is an important result given that these numeric features can be simply estimated from the corpus without recourse to taxonomic information. Pairing the WordNet frequency feature $f_{wn}(n_1,n_2)$ with any other feature yields results comparable to using solely this feature. Only a .5% increase over 72.0% is achieved when $f_{wn}(n_1,n_2)$ is combined with the frequency of the head noun $f(n_2)$. Note that WordNet outperforms Roget's thesaurus even though both dictionaries contain taxonomic information. In fact, pairing Roget with any numerical feature other than WordNet yields a lower accuracy than pairing $P(\mathrm{H}|n_1)$ with $f(n_2)$ (see Table 6.9).

Table 6.10 reports classification accuracy using triples of the six numeric features. When three non-taxonomic features are used, the performance fluctuates between 59.5% (when $P(\mathrm{M}|n_2)$, $f(n_1)$, and $f(n_2)$ are combined) and 67.5% (for $P(\mathrm{H}|n_1)$, $P(\mathrm{M}|n_2)$, and $f(n_2)$). The combination of WordNet with any two non-taxonomic features is steadily good, without however outperforming WordNet alone. The combination of Roget with two non-taxonomic features improves accuracy (69.9% for $f_{ro}(n_1,n_2)$, $P(\mathrm{M}|n_2)$, and $f(n_1)$) over its combination with a single non-taxonomic feature (67.5% for $f_{ro}(n_1,n_2)$, and $P(\mathrm{M}|n_2)$). The best result is achieved

Table 6.10: Decision trees with three numerical features

Features	Train (%)	Test (%)
D	56.9 ± 1.97	54.5 ± 3.53
$P(\mathrm{H}\|n_1)$, $P(\mathrm{M}\|n_2)$, $f(n_1)$	62.9 ± 1.92	63.0 ± 3.42
$P(\mathrm{H}\|n_1)$, $P(\mathrm{M}\|n_2)$, $f(n_2)$	64.2 ± 1.99	67.5 ± 3.32
$P(\mathrm{H}\|n_1)$, $f(n_1)$, $f(n_2)$	61.1 ± 1.95	62.0 ± 3.44
$P(\mathrm{M}\|n_2)$, $f(n_1)$, $f(n_2)$	61.2 ± 1.94	59.5 ± 3.48
$f_{wn}(n_1,n_2)$, $P(\mathrm{H}\|n_1)$, $P(\mathrm{M}\|n_2)$	66.4 ± 1.88	72.0 ± 3.18
$f_{wn}(n_1,n_2)$, $P(\mathrm{M}\|n_2)$, $f(n_1)$	66.4 ± 1.88	72.0 ± 3.18
$f_{wn}(n_1,n_2)$, $P(\mathrm{M}\|n_2)$, $f(n_2)$	67.4 ± 1.87	68.5 ± 3.29
$f_{wn}(n_1,n_2)$, $P(\mathrm{H}\|n_1)$, $f(n_1)$	66.4 ± 1.88	72.0 ± 3.18
$f_{wn}(n_1,n_2)$, $P(\mathrm{H}\|n_1)$, $f(n_2)$	66.4 ± 1.88	72.0 ± 3.18
$f_{wn}(n_1,n_2)$, $f(n_1)$, $f(n_2)$	66.6 ± 1.88	69.5 ± 3.26
$f_{ro}(n_1,n_2)$, $P(\mathrm{H}\|n_1)$, $P(\mathrm{M}\|n_2)$	63.7 ± 1.92	67.0 ± 3.33
$f_{ro}(n_1,n_2)$, $P(\mathrm{M}\|n_2)$, $f(n_1)$	62.6 ± 1.93	69.9 ± 3.28
$f_{ro}(n_1,n_2)$, $P(\mathrm{M}\|n_2)$, $f(n_2)$	63.2 ± 1.93	69.0 ± 3.28
$f_{ro}(n_1,n_2)$, $P(\mathrm{H}\|n_1)$, $f(n_1)$	62.9 ± 1.92	67.0 ± 3.33
$f_{ro}(n_1,n_2)$, $P(\mathrm{H}\|n_1)$, $f(n_2)$	65.1 ± 1.90	68.0 ± 3.31
$f_{ro}(n_1,n_2)$, $f(n_1)$, $f(n_2)$	62.9 ± 1.92	64.5 ± 3.39
$f_{wn}(n_1,n_2)$, $f_{ro}(n_1,n_2)$, $P(\mathrm{H}\|n_1)$	67.2 ± 1.87	74.0 ± 3.11
$f_{wn}(n_1,n_2)$, $f_{ro}(n_1,n_2)$, $P(\mathrm{M}\|n_2)$	66.4 ± 1.88	71.5 ± 3.20
$f_{wn}(n_1,n_2)$, $f_{ro}(n_1,n_2)$, $f(n_1)$	68.0 ± 1.86	68.5 ± 3.29
$f_{wn}(n_1,n_2)$, $f_{ro}(n_1,n_2)$, $f(n_2)$	67.4 ± 1.86	70.0 ± 3.25

for the combination of $f_{wn}(n_1,n_2)$, $f_{ro}(n_1,n_2)$, and $P(\mathrm{H}|n_1)$. These three features achieve an accuracy of 74.0% which outperforms WordNet alone although the difference is not statistically significant ($p = .6$).

Accuracy on the classification task does not reach 74.0% when quadruples of features are used (see Table 6.11). In fact, the highest performance is 72.0% for all combinations of Roget and WordNet with any two non-taxonomic features and for WordNet when combined respectively with $P(\mathrm{H}|n_1)$, $P(\mathrm{M}|n_2)$, and $f(n_1)$ and $P(\mathrm{M}|n_2)$, $f(n_1)$, and $f(n_2)$. The combination of all four non-taxonomic features ($P(\mathrm{H}|n_1)$, $P(\mathrm{M}|n_2)$, $f(n_1)$, and $f(n_2)$) yields a 5.5% decrease in performance over the combination of the features $P(\mathrm{H}|n_1)$ and $f(n_2)$ (compare Table 6.9). When quintuples of features are used, the highest performance (72.5%) is achieved for $f_{wn}(n_1,n_2)$, $f_{ro}(n_1,n_2)$, $P(\mathrm{M}|n_2)$, $f(n_1)$, and $f(n_2)$. Note that the lowest performances are achieved when a taxonomic feature is combined with four non-taxonomic ones (66.0% when WordNet is combined with $P(\mathrm{M}|n_2)$, $P(\mathrm{H}|n_1)$, $f(n_1)$, and $f(n_2)$ and 61.5% when Roget is combined with the same features). Combination of all six features (see Table 6.12) achieves a performance of 71.5% without outperforming WordNet alone (see Table 6.8).

In sum, the best results (74.0%) are achieved with the combination of two taxonomic features ($f_{wn}(n_1,n_2)$ and $f_{ro}(n_1,n_2)$) and one non-taxonomic feature $P(\mathrm{H}|n_1)$. Without using taxonomic knowledge the combination of $P(\mathrm{H}|n_1)$ and $f(n_2)$ yields the best performance (68.0%). Both performances are significantly better than the baseline of 54.5% ($p < .01$). A

Table 6.11: Decision trees with four numerical features

Features	Train (%)	Test (%)
D	56.9 ± 1.97	54.5 ±3.53
$P(\mathrm{H}\|n_1)$, $P(\mathrm{M}\|n_2)$, $f(n_1)$, $f(n_2)$	65.1 ± 1.90	62.5 ± 3.43
$f_{wn}(n_1,n_2)$, $P(\mathrm{H}\|n_1)$, $P(\mathrm{M}\|n_2)$, $f(n_1)$	66.4 ± 1.88	72.0 ± 3.18
$f_{wn}(n_1,n_2)$, $P(\mathrm{H}\|n_1)$, $P(\mathrm{M}\|n_2)$, $f(n_2)$	68.8 ± 1.85	67.0 ± 3.33
$f_{wn}(n_1,n_2)$, $P(\mathrm{H}\|n_1)$, $f(n_1)$, $f(n_2)$	68.6 ± 1.85	67.0 ± 3.33
$f_{wn}(n_1,n_2)$, $P(\mathrm{M}\|n_2)$, $f(n_1)$, $f(n_2)$	66.4 ± 1.88	72.0 ± 3.18
$f_{ro}(n_1,n_2)$, $P(\mathrm{H}\|n_1)$, $P(\mathrm{M}\|n_2)$, $f(n_1)$	63.9 ± 1.91	69.0 ± 3.28
$f_{ro}(n_1,n_2)$, $P(\mathrm{H}\|n_1)$, $P(\mathrm{M}\|n_2)$, $f(n_2)$	66.1 ± 1.89	70.0 ± 3.25
$f_{ro}(n_1,n_2)$, $P(\mathrm{H}\|n_1)$, $f(n_1)$, $f(n_2)$	63.1 ± 1.92	64.0 ± 3.40
$f_{ro}(n_1,n_2)$, $P(\mathrm{M}\|n_2)$, $f(n_1)$, $f(n_2)$	62.8 ± 1.93	69.5 ± 3.26
$f_{wn}(n_1,n_2)$, $f_{ro}(n_1,n_2)$, $P(\mathrm{H}\|n_1)$, $P(\mathrm{M}\|n_2)$	66.4 ± 1.88	72.0 ± 3.18
$f_{wn}(n_1,n_2)$, $f_{ro}(n_1,n_2)$, $P(\mathrm{H}\|n_1)$, $f(n_1)$	66.4 ± 1.88	72.0 ± 3.18
$f_{wn}(n_1,n_2)$, $f_{ro}(n_1,n_2)$, $P(\mathrm{H}\|n_1)$, $f(n_2)$	67.4 ± 1.88	72.0 ± 3.18
$f_{wn}(n_1,n_2)$, $f_{ro}(n_1,n_2)$, $P(\mathrm{M}\|n_2)$, $f(n_1)$	66.2 ± 1.88	72.0 ± 3.18
$f_{wn}(n_1,n_2)$, $f_{ro}(n_1,n_2)$, $P(\mathrm{M}\|n_2)$, $f(n_2)$	66.4 ± 1.88	72.0 ± 3.18
$f_{wn}(n_1,n_2)$, $f_{ro}(n_1,n_2)$, $f(n_1)$, $f(n_2)$	66.4 ± 1.88	72.0 ± 3.18

Table 6.12: Decision trees with five and six numerical features

Features	Train (%)	Test (%)
D	56.9 ± 1.97	54.5 ± 3.53
$f_{wn}(n_1,n_2)$, $P(\mathrm{M}\|n_2)$, $P(\mathrm{H}\|n_1)$, $f(n_1)$, $f(n_2)$	68.9 ± 1.84	66.0 ± 3.36
$f_{ro}(n_1,n_2)$, $P(\mathrm{M}\|n_2)$, $P(\mathrm{H}\|n_1)$, $f(n_1)$, $f(n_2)$	61.5 ± 1.94	61.5 ± 3.45
$f_{wn}(n_1,n_2)$, $f_{ro}(n_1,n_2)$, $P(\mathrm{M}\|n_2)$, $f(n_1)$, $f(n_2)$	67.2 ± 1.87	72.5 ± 3.17
$f_{wn}(n_1,n_2)$, $f_{ro}(n_1,n_2)$, $P(\mathrm{H}\|n_1)$, $f(n_1)$, $f(n_2)$	69.4 ± 1.84	68.5 ± 3.29
$f_{wn}(n_1,n_2)$, $f_{ro}(n_1,n_2)$, $P(\mathrm{M}\|n_2)$, $P(\mathrm{H}\|n_1)$, $f(n_1)$	67.5 ± 1.87	69.0 ± 3.28
$f_{wn}(n_1,n_2)$, $f_{ro}(n_1,n_2)$, $P(\mathrm{M}\|n_2)$, $P(\mathrm{H}\|n_1)$, $f(n_2)$	66.4 ± 1.88	72.0 ± 3.18
$f_{wn}(n_1,n_2)$, $f_{ro}(n_1,n_2)$, $P(\mathrm{H}\|n_1)$, $P(\mathrm{M}\|n_2)$, $f(n_1)$, $f(n_2)$	67.6 ± 1.87	71.5 ± 3.20

general observation is that performance increases when taxonomic knowledge is used, even though its use relies on unrealistic assumptions about concept combinations (see Section 6.9.1). The use of the WordNet taxonomy consistently outperforms Roget's thesaurus. This fact may be due to the size of the taxonomies. WordNet contains twice as many noun entries as Roget (47,302 versus 20,448). Another explanation might be that Roget's thesaurus is too coarse-grained a taxonomy for the task at hand (Roget's taxonomy contains 1,043 concepts, whereas WordNet contains 4,795).

An important question is whether context alone can provide cues for classifying the candidate compounds. Recall from Section 6.9.1 that context was encoded as parts of speech. We evaluated the influence of context by varying both the position and the size of the window of words (i.e., parts of speech) surrounding the candidate compound. We varied the window size parameter between one and four words before and after the candidate compounds. We use symbols l and r for left and right context, respectively and numbers to denote the size of the

Table 6.13: Influence of asymmetric context

Features	Train (%)	Test (%)
D	56.9 ± 1.97	54.5 ± 3.53
$l = 0,\ r = 1$	70.8 ± 1.81	68.5 ± 3.29
$l = 0,\ r = 2$	69.6 ± 1.83	68.5 ± 3.29
$l = 0,\ r = 3$	71.8 ± 1.79	66.0 ± 3.36
$l = 0,\ r = 4$	69.6 ± 1.83	69.5 ± 3.26
$l = 1,\ r = 2$	71.0 ± 1.81	70.0 ± 3.25
$l = 1,\ r = 3$	71.2 ± 3.33	67.0 ± 3.33
$l = 1,\ r = 4$	71.2 ± 3.33	67.0 ± 3.33
$l = 2,\ r = 3$	71.0 ± 1.81	67.5 ± 3.32
$l = 2,\ r = 4$	70.6 ± 1.82	64.0 ± 3.40
$l = 3,\ r = 4$	69.9 ± 1.83	64.0 ± 3.40

Table 6.14: Influence of asymmetric context

Features	Train (%)	Test (%)
D	56.9 ± 1.97	54.5 ± 3.53
$r = 0,\ l = 1$	69.6 ± 1.83	73.0 ± 3.13
$r = 0,\ l = 2$	66.7 ± 1.88	65.0 ± 3.38
$r = 0,\ l = 3$	68.9 ± 1.84	69.0 ± 3.28
$r = 0,\ l = 4$	68.6 ± 1.85	69.5 ± 3.26
$r = 1,\ l = 2$	69.9 ± 1.83	70.0 ± 3.25
$r = 1,\ l = 3$	69.9 ± 1.83	70.0 ± 3.25
$r = 1,\ l = 4$	70.1 ± 1.82	65.0 ± 3.38
$r = 2,\ l = 3$	70.7 ± 1.81	66.0 ± 3.36
$r = 2,\ l = 4$	70.4 ± 1.82	66.0 ± 3.36
$r = 3,\ l = 4$	70.5 ± 1.82	64.0 ± 3.40

Table 6.15: Influence of symmetric context

Features	Train (%)	Test (%)
D	56.9 ± 1.97	54.5 ± 3.53
$l = 1,\ r = 1$	71.6 ± 1.80	67.0 ± 3.33
$l = 2,\ r = 2$	71.8 ± 1.81	66.0 ± 3.36
$l = 3,\ r = 3$	72.3 ± 1.78	66.0 ± 3.36
$l = 4,\ r = 4$	70.5 ± 1.82	64.0 ± 3.40

window surrounding the candidate compound. For example, $l = 2$, $r = 4$ represents a window of two words to left and four words to the right of the candidate noun-noun sequence.

Tables 6.13 and 6.14 explore the influence of asymmetric context in the classification of the candidate compounds, whereas Table 6.15 shows the influence of symmetric context. The best performance (73.0%) is achieved with the minimal window of one word to the left of the candidate compound ($r = 0$, $l = 1$, see Table 6.14). Good performance is also achieved with larger asymmetric contexts: a context of one word to the left and two words to the right ($l = 1$, $r = 2$) achieves an accuracy of 70.0% (see Table 6.13). The same accuracy is achieved with two or three words to the left and one word to the right (see Table 6.14). Performance decreases with large context windows, either asymmetric (see $l = 3$, $r = 4$ in Table 6.13 and $r = 3$, $l = 4$ in Table 6.14) or symmetric (see $l = 4$, $r = 4$ in Table 6.15). A smaller context captures local syntactic dependencies such as the fact that compound nouns are typically preceded by determiners, verbs, or adjectives and succeeded by verbs, prepositions, or function words (e.g., *and*, *or*, *as*). When a larger context is taken into account the local grammatico-syntactic dependencies are somewhat lost since more variety is introduced in the vectors representing the context surrounding the candidate compound and generalizations do not emerge as easily.

Note that context alone is sufficient to classify candidate compounds with precision

Table 6.16: Agreement between numerical and categorial features

	$f(n_1)$	$f(n_2)$	$P(H\|n_1)$	$P(M\|n_2)$	$f_{wn}(n_1,n_2)$	$f_{ro}(n_1,n_2)$
$f(n_2)$	−.04					
$P(H\|n_1)$	.56	.10				
$P(M\|n_2)$	−.01	.74	.09			
$f_{wn}(n_1,n_2)$	.24	.24	.18	.34		
$f_{ro}(n_1,n_2)$	.40	.33	.28	.34	.41	
$l=1, r=0$	.01	.27	.10	.30	.19	.17

comparable to numeric features. Our results suggest that grammatical knowledge (i.e., parts of speech) and syntactic dependencies provide enough information to distinguish compounds from non-compounds. This is an important result given that our best numerical predictor (i.e., $f_{wn}(n_1,n_2)$, $f_{ro}(n_1,n_2)$, and $P(H|n_1)$) relies heavily on taxonomic information. The contextual features are straightforward to obtain—all we need is a concordance of the candidate compound annotated with parts of speech.

Before exploring how the combination of categorial and numerical features performs at classifying candidate compounds, we examine the extent to which categorial and numerical features agree in their classifications. We measure the different features' agreement on the classification task using the Kappa coefficient (see Section 2.5.1). We calculate Kappa for all pairwise combinations of the five numeric features ($f(n_1)$, $f(n_2)$, $P(H|n_1)$, $P(M|n_2)$, $f_{wn}(n_1,n_2)$, $f_{ro}(n_1,n_2)$) and the best contextual feature ($r=0$, $l=1$). The results are given in Table 6.16 The highest agreement is observed for $f(n_2)$ and $P(M|n_2)$ ($K=.74$). Generally speaking, agreement between taxonomic and simple corpus-based features is low ranging from .18 to .34 (when $f_{wn}(n_1,n_2)$ is paired with $P(H|n_1)$ and $P(M|n_2)$, respectively) and from .28 to .40 (when $f_{ro}(n_1,n_2)$ is paired with $P(H|n_1)$ and $f(n_1)$, respectively). Agreement between the two taxonomic features ($f_{wn}(n_1,n_2)$ and $f_{ro}(n_1,n_2)$) is also low ($K=.41$) as well as agreement between the best contextual feature ($l=1$, $r=0$) and any other feature (K ranges from .01 to .17). These results indicate that there is little overlap in the classifications provided by the individual features suggesting that they are complementary. This is not surprising given that these features represent different types of distributional information. Taxonomic features capture lexical semantic information, whereas contextual features capture syntactic dependencies.

Therefore, we further examine how the combination of categorial and numerical features performs at classifying candidate compounds. Table 6.17 shows how accuracy varies when the best context is paired with the best taxonomic and non-taxonomic features. Consideration of the context surrounding the candidate compound generally increases performance. The combination of contextual features with taxonomic features outperforms the individual learning methods achieving a significant increase of 20.5% ($p<.01$) over the baseline of always classifying a noun-noun sequence as a compound. The combination of contextual and non-

Table 6.17: Combination of numerical and categorial features

Features	Train (%)	Test (%)
D	56.9 ± 1.97	54.5 ± 3.53
$r = 1,\ l = 0,\ f_{wn}(n_1,n_2)$	71.3 ± 1.80	75.0 ± 3.07
$r = 0,\ l = 1,\ f_{wn}(n_1,n_2),\ f_{ro}(n_1,n_2),\ P(\mathrm{H}\|n_1)$	71.0 ± 1.81	74.5 ± 3.07
$r = 1,\ l = 0,\ P(\mathrm{H}\|n_1),\ f(n_2)$	67.2 ± 1.87	71.0 ± 3.22

Table 6.18: Performance on classifying compounds on test data

Features	Compound (%)	Non-Compound (%)
$f_{wn}(c_1,c_2)$	81.7	60.4
$f_{wn}(c_1,c_2),\ f_{ro}(c_1,c_2)\ P(\mathrm{H}\|n_1)$	90.8	53.8
$r = 0, l = 1$	81.7	62.6
$r = 0, l = 1, f_{wn}(c_1,c_2)$	85.3	62.6

taxonomic features (i.e., $P(\mathrm{H}|n_1)$ and $f(n_2)$) achieves an accuracy of 71.0% (see Table 6.17). The result shows that simple features which can be easily retrieved and estimated from the corpus contain enough information to capture generalizations about the behavior of compounds.

Table 6.18 shows how the different feature combinations compare on classifying compounds and non-compounds. Here we consider only the best individual numerical feature (i.e., $f_{wn}(n_1,n_2)$), the best combination of numerical features (i.e., $f_{wn}(n_1,n_2)$, $f_{ro}(n_1,n_2)$, $P(H|n_1)$), the best contextual features (i.e., $r = 0$, $l = 1$), and the best combination of categorial and numerical features (i.e., $r = 0$, $l = 1$, $f_{wn}(n_1,n_2)$, $f_{ro}(n_1,n_2)$, and $P(\mathrm{H}|n_1)$). Note that all features perform well at classifying valid compounds. Classification of non-compounds is a considerably harder task (see the discussion in Section 6.10) for which satisfactory performance is achieved when taxonomic or contextual information is taken into account (60.4% for $f_{wn}(n_1,n_2)$ and 62.6% for $r = 0$, $l = 1$, see Table 6.18).

6.10. Discussion

We presented a method that distinguishes compounds from non-compounds by taking into account the context within which the candidate compound occurs, taxonomic information capturing the likelihood of two concepts to form a compound, and numerical features encoding the likelihood of a given word as a compound head or a compound modifier. We used decision tree learning to assess their contribution and showed that, even without taking taxonomic information into account, a significant increase of 16.5% ($p < .01$) can be achieved over a baseline of 54.5% (see Table 6.17). We further showed that contextual and taxonomic features are individually very good predictors and that context performs best at distinguishing non-compounds (see Tables 6.8, 6.14, and 6.18).

Our results are encouraging considering the simplicity of the features we took into ac-

count for our machine learning experiments. The task of deciding whether two nouns form a compound or not crucially depends on a variety of factors such as world-knowledge, the situation at hand, and the speaker's and hearer's communicative goals, none of which are directly represented by our features. Consider for example the sequence *organization time* which is classified as a valid compound by our features (see the feature combinations in Table 6.17). Out of context *organization time* can be interpreted as "time it takes to organize something", "time during which something is organized" or "time for organization". Sentence (6.21) gives the context within which *organization time* is attested. In this particular example *organization time* is not a compound, even if in the general case it could be. Context cannot provide clues for classifying this candidate compound correctly, one would need to know the subcategorization preferences of the verb *give* (i.e., that it takes a double object), and that *organization* does not denote an activity in this context but rather an administration body.

(6.21) He implants ideas, gives clues, prompts proposals, and avoids committing himself publicly until the last moment. This gives the [organization$_N$ time$_N$] to absorb the idea, to build consensus and to reduce resistance.

The experiments presented in this chapter provide further support for the surface cueing approach advocated throughout this thesis. In Chapter 3 we acquired verb frames characteristic of diathesis alternations relying on Levin's generalizations about the meaning of verbs and their syntactic behavior. Our acquisition experiments were guided by surface syntactic, grammatical, and semantic cues. In this chapter we also explored the contribution of such cues for a very productive linguistic phenomenon. Although diathesis alternations are exhibited by several verbs, it is not the case that any verb can exhibit any alternation. In fact, the meaning of the verb and its syntactic frames constrain the type of alternations it can undergo. Any sequence of two nouns can be a potential compound given the right context. This means that not only frequent noun combinations are potential compounds but also rare ones. A surface approach which looks for consecutive nouns in the corpus is only sufficient for frequent compounds (see Experiments 10–14). For the case of rare compounds, we showed that the combination of several surface cues yields satisfactory results. As in Chapter 3 we combined surface syntactic cues with semantic cues.

Our experiments revealed that syntactic cues such as the frequency of the compound head, the likelihood of a word as a modifier, or the context surrounding a candidate compound perform almost as well as cues that are estimated on the basis of existing taxonomies such as WordNet (e.g., the frequency of the concepts representing the candidate compound). We further demonstrated that the surface cueing approach can overcome the problem of sparse data which is closely related to the productivity of compounding. In particular, by exploiting information about frequent compounds or frequent contexts (which can be easily retrieved from the corpus) we can "recreate" evidence about the likelihood of two nouns to form a valid compound.

6.11. Related Work

Several types of statistical tests have been proposed in the literature for the discovery of terms from corpora. Co-occurrence frequency has been used in a variety of studies and has been shown to perform surprisingly well given its simplicity. Justeson and Katz (1995c) show that a criterion for terminology identification is simple repetition—any NP with frequency greater or equal to two is a candidate term. A variety of statistical scores assess whether two words co-occur together more often than chance. Church and Hanks (1990) have used the t test for the discovery of collocations (e.g., *strong tea*, *strong support*). Smadja (1992) discovers collocations by taking into account the mean and variance of the frequency with which a target word co-occurs with its surrounding words within a context of ten words. Both methods could be easily adapted for the acquisition of terms and compound nouns. Other statistical tests include χ^2 and the log-likelihood ratio.[4] Dunning (1993) has shown that the log-likelihood ratio is more appropriate for the discovery of terms, especially for small frequencies. Mutual information (Church and Hanks 1990) is also a popular statistic for the discovery of lexically associated terms. A well-known property of mutual information is its tendency to overestimate rare events.

Daille (1996) performed a study where she compared the appropriateness of several statistical scores for the task of terminology discovery. The statistical scores included the standardly used co-occurrence frequency, mutual information, the log-likelihood ratio, as well as the less well-known cubic association ratio (a variant of mutual information which gives less weight to rare events), the Φ^2 co-efficient, the Yule co-efficient, the Kulczinsky co-efficient, the Fager and McGowan co-efficient (see Daille 1996 for a definition of these scores). Daille (1996) found that co-occurrence frequency and the log-likelihood ratio were the most effective scores for the identification of valid terms.

A common feature of the approaches described so far is the lack of detailed knowledge with regard to the acquisition task. In some cases part-of-speech or syntactic information is taken into account (e.g., Justeson and Katz 1995c, Daille 1996) without however exploiting pre-existing terminological dictionaries. Jacquemin (1996) acquires new terms by observing the variants of existing terminological lists in the corpus. During the acquisition not only new terms are discovered but also conceptual links between semantically related terms.

The automatic acquisition of compound nouns (as opposed to terms) from unrestricted wide-coverage text has not received much attention in the literature. Lauer's (1995) study was conducted on a corpus exhibiting a uniform register and was furthermore biased in favor of syntactically unambiguous nouns. It cannot therefore be considered representative of part-of-

[4] Fisher's exact test can be also used to discover if two words are associated more often than chance (Pedersen, Kayaalp, and Bruce 1996). The test is appropriate for small counts, however it is cumbersome to compute, especially for large data samples. The test has not been applied to the discovery of terms or compounds and it is unclear whether its performance is substantially better than the χ^2 test or the log-likelihood ratio (see Weeber, Vos, and Baayen 2000 for a comparison between the log-likelihood ratio and Fisher's exact test).

speech tagged domain independent text. Our approach focused on compounds for which very little evidence is found in the corpus. None of the approaches described above are appropriate for this task since they rely on the assumption that two words form a term when they co-occur together often enough or more often than chance. We explored the contribution of various non-taxonomic and taxonomic features to the acquisition task. The non-taxonomic features were estimated from the corpus without recourse to external knowledge sources, by exploiting the distributional properties of established (i.e., frequent) compounds.

Our approach is conceptually close to Jacquemin (1996): in both cases a list of terms is used for the acquisition task. The crucial difference is that our approach does not presuppose the availability of a list of established terms external to the corpus for the acquisition to take place. We rely solely on the corpus for the discovery of the reliable compounds (noun-noun sequences with CoocF$\geq$5) from which our numerical features are estimated. Another difference is that we discover novel compounds, whereas Jacquemin's (1996) method can only discover variants of already existing terms. Our taxonomic features were estimated by combining corpus evidence with domain independent taxonomic information. Taxonomic information has been used for a variety of word sense disambiguation tasks (e.g., Yarowsky 1992, Voorhees 1993, Agirre and Rigau 1996), for the acquisition of selectional restrictions (e.g., Resnik 1993, Ciaramita and Johnson 2000) and hyponyms (Hearst 1992), and for syntactic disambiguation (Brill and Resnik 1994). We demonstrate that taxonomic information yields satisfactory results when applied to an additional lexical phenomenon, i.e., the acquisition of compound nouns.

A considerable body of research has concentrated on the identification of features for the automatic classification of linguistic phenomena. Siegel and McKeown (1994) apply decision tree learning to the disambiguation of discourse clues (e.g., the word *say* is ambiguous between a discourse sense where it means "for example" and sentential sense where it means "express"). Passonneau and Litman (1997) use linguistic features for the task of discourse segmentation. Siegel (1999) uses corpus-based linguistic indicators to classify verbal aspectual preferences (e.g., events or states). Merlo and Stevenson (1999) use grammatical features for the classification of verbs into lexical semantic classes (e.g., unergative, unaccusative, and object-drop). Hatzivassiloglou and McKeown (1995a) and Hatzivassiloglou and McKeown (1997) present a method which exploits several linguistic features in order to identify the semantic orientation of adjectives (e.g., *intelligent* receives a positive orientation, whereas *stupid* receives a negative orientation, see Chapter 5 for details). Hatzivassiloglou and Wiebe (2000) exploit numeric features for the classification of gradable and non-gradable adjectives (see Section 5.7).

The use of machine-learning for the identification of compound nouns is novel to our knowledge. Research in the acquisition of terms from corpora has focussed on word sequences whose recurrent occurrence in a corpus is indicative of their terminological status, and therefore has not addressed the issue of discovering terms when data is sparse. This becomes more

apparent in the case of noun compounding which is an extremely productive process (Downing 1977). Is a co-occurrence frequency count of one merely the result of insufficient evidence, or is it a reflection of a tagging or parsing mistake? We have shown that data-sparseness can be overcome by taking into account a large number of linguistic features. We have evaluated the contribution of these features using standard machine learning techniques and shown that information inherent in the corpus can make up for the lack of distributional evidence in the case of hapaxes.

6.12. Summary

In this chapter we focused on the acquisition of compounds from wide coverage text. We showed that a simple heuristic which looks for consecutive nouns in the BNC (see Section 6.2) results in substantially lower accuracy (i.e., 71.0%) than previously reported in the literature (i.e., 97.9%, Lauer 1995). Statistical scores such as co-occurrence frequency and log-likelihood ratio allow us to establish thresholds above which we can consider candidate noun-noun sequences as compounds with high degrees of accuracy (see Table 6.4). This means, however, that only a small fraction of the compounds attested in the corpus is acquired.

Experiment 14 revealed that hapaxes are the majority of the candidate compounds extracted from the corpus (see Section 6.7). 57.7% of these noun-noun sequences are valid compounds. A large number of these compounds are novel (i.e., created on the spot to satisfy the speaker's communicative needs) and cannot be distinguished from nonce terms solely on the basis of their distributional properties. We presented a method that distinguishes rare compounds from non-compounds by taking into account the context within which the candidate compound occurs, taxonomic information capturing the likelihood of two concepts to form a compound, and numerical features encoding the likelihood of a given word as a head or a modifier. We used decision tree learning to assess their contribution and showed that, even without taking taxonomic information into account, a significant increase of 16.5% ($p < .01$) can be achieved over a baseline of 54.5% (see Table 6.17). We further showed that contextual and taxonomic features are individually very good predictors (see Tables 6.8 and 6.14) and that context performs best at distinguishing non-compounds (see Table 6.18).

In the following chapter we turn to the interpretation of nominalizations (a particular type of compounds whose heads are derived from a verb and whose modifiers are arguments of this verb) and present a simple probabilistic model that can be used to infer the argument relation between a nominalized head and its modifier. The interpretation of compounds also runs into severe data sparseness: not only is the argument relation between the head and its modifier unattested in the corpus, but an approximation which maps the head to its underlying verb also provides insufficient evidence. We address the interpretation problem by experimenting with smoothing techniques that exploit distributional and taxonomic information to "recreate"

the frequency of word combinations unattested in the corpus. As in this chapter, the different information sources are combined using decision tree learning.

Chapter 7

A Probabilistic Model of Nominalizations

In this chapter we present a probabilistic model for the interpretation of nominalizations, a particular class of compound nouns whose head noun is derived from a verb and whose modifier is interpreted as the argument of this verb. We show how the probabilistic model presented in Chapters 4 and 5 can be used to disambiguate nominalizations by inferring the argument relation (e.g., subject or object) between the modifier and its nominalized head. Similarly to Chapter 5 we use the corpus as the inventory of the argument relations. Although the degree of ambiguity exhibited by nominalizations is less than that exhibited by polysemous verbs and adjectives (in the case of nominalizations we have a binary choice between a subject or object relation), the task is less straightforward than it seems since the estimation of the model parameters runs into severe data sparseness problems: the semantic relation between the head and its modifier is not attested in the corpus and even an approximation which maps the compound head to its underlying verb provides insufficient evidence. As in the previous chapter, we overcome data sparseness using surface cueing (i.e., partial parsing) together with a variety of smoothing techniques which "recreate" the missing evidence by exploiting different types of information using decision tree learning for their combination. This chapter provides further support for the proposed approach showing how the ambiguity of nominalizations can be reduced even in cases of data sparseness by taking advantage of information present in the corpus.

7.1. Introduction

The automatic interpretation of compound nouns has been a long-standing unsolved problem for NLP. Compound nouns in English have three basic properties which pose difficulties for their interpretation: (a) the compounding process is extremely productive (this means that a

hypothetical system would have to interprete previously unseen instances), (b) the semantic relationship between the compound head and its modifier is implicit (this means that it cannot be easily recovered from syntactic or morphological analysis), and (c) the interpretation can be influenced by a variety of contextual and pragmatic factors.

A considerable amount of effort has gone into specifying the set of semantic relations that hold between a compound head and its modifier (Finin 1980; Isabelle 1984; Levi 1978; Warren 1978). Levi (1978), for example, distinguishes two types of compound nouns: (a) compounds consisting of two nouns which are related by one of nine recoverably deletable predicates (e.g., CAUSE relates *onion tears*, FOR relates *pet spray*, see the examples in (7.1)) and (b) nominalizations, i.e., compounds whose heads are nominalizations, i.e., nouns derived from a verb, and their modifiers are interpreted as arguments of the related verb (e.g., a *car lover* loves cars, see the examples in (7.2)–(7.4)). The prenominal modifier can be either a noun or an adjective (see the examples in (7.2)). The nominalized verb can either take a subject (see (7.3a)), a direct object (see (7.3b)) or a prepositional object (see (7.3c)).

(7.1)	a.	onion tears	CAUSE
	b.	vegetable soup	HAVE
	c.	music box	MAKE
	d.	steam iron	USE
	e.	pine tree	BE
	f.	night flight	IN
	g.	pet spray	FOR
	h.	peanut butter	FROM
	i.	abortion problem	ABOUT
(7.2)	a.	parental refusal	SUBJ
	b.	cardiac massage	OBJ
	c.	heart massage	OBJ
	d.	sound synthesizer	OBJ
(7.3)	a.	child behaviour	SUBJ
	b.	car lover	OBJ
	c.	soccer competition	OBJ
(7.4)	a.	government promotion	SUBJ|OBJ
	b.	satellite observation	SUBJ|OBJ

Besides Levi (1978), a fair number of researchers (e.g., Warren 1978, Finin 1980, Isabelle 1984, and Leonard 1984) agree that there is a limited number of regularly recurring relations between a compound head and its modifier. There is far less agreement when it comes to the type and number of these relations. The relations vary from Levi's (1978) recoverably deletable predicates to Warren's (1978) paraphrases, and Finin's (1980) role nominals. Leonard

(1984) proposes eight relations, Warren (1978) proposes six basic relations, whereas the number of relations proposed by Finin (1980) are potentially infinite.

The attempt to restrict the semantic relations between the compound head and its modifier to a prespecified number and type has been criticized by Downing (1977), who has shown (through a series of psycholinguistic experiments) that the underlying relations can be influenced by a variety of pragmatic factors and cannot be therefore presumed to be easily enumerable. Sparck Jones (1983: 4) further notes "that observations about the semantic relation holding between the compound head and its modifier can only be remarks about tendencies and not about absolutes". Consider for instance the compound *onion tears* (see (7.1a)). The relationship CAUSE is one of the possible interpretations the compound may receive. One could easily imagine a context where the tears are FOR or ABOUT the onion. Consider example (6.5a), repeated here as (7.5a), where *apple-juice seat* refers to the situation where someone is instructed to sit in a seat in front of which a glass of apple-juice has been placed. Given this particular state of affairs, none of the relations in (7.1) can be used to successfully interpret *apple-juice seat.* Such considerations have led Selkirk (1982) to claim that only nominalizations are amenable to linguistic characterization, leaving all other compounds to be explained by pragmatics or discourse. A similar approach is put forward by Hobbs, Stickel, Appelt, and Martin (1993) for all types of compounds, including nominalizations: any two nouns can be combined, and the relation between these nouns is entirely underspecified, to be resolved pragmatically.

(7.5) a. A friend of mine was once instructed to sit in the [apple-juice$_N$ seat$_N$]. (Downing 1977: 818)

b. By the end of the 1920s, [government$_N$ promotion$_N$] of agricultural development in Niger was limited, consisting mainly of crop trials and model sheep and ostrich farms.

Less controversy arises with regard to nominalizations, perhaps due to the small number of allowable relations. Most approaches follow Levi (1978) in distinguishing nominalizations as a separate class of compounds, the exception being Finin (1980) who claims that most compounds are nominalizations, even in cases where the head noun is not morphologically derived from a verb (see the examples in (7.1)). Under Finin's (1980) analysis the head *book* in the compound *recipe book* is a role nominal, i.e., a noun which refers to a particular thematic role of another concept. This means that *book* refers to the object role of *write* which is filled by *recipe.* However, it is not clear how the implicit verb is to be recovered or why *write* is more appropriate than *read* in this example.

Despite the small number of relations between the nominalized head and its modifier, the interpretation of nominalizations can readily change in different contexts. In some cases, the relation of the modifier and the nominalized verb (e.g., subject or object) can be predicted either from the subcategorization properties of the verb or from the semantics of the nominalization suffix of the head noun. Consider (7.3a) for example. Here *child* can only be the subject

of *behaviour*, since the verb *behave* is intransitive. In (7.3b) the agentive suffix *-er* of the head noun *lover* indicates that the modifier *car* is the object of the verb *love*. In other cases, the relation of the modifier and the head noun is genuinely ambiguous. Out of context the compounds *government promotion* and *satellite observation* (see examples (7.4)) can receive either a subject or an object interpretation. One might argue that the preferred analysis for *government promotion* is "government that is promoted by someone". However, this interpretation can be easily overridden in context as shown in example (7.5b) taken from the BNC: here it is the government that is doing the promotion.

The automatic interpretation of compound nouns poses a challenge for empirical approaches since the relations between a head and its modifier are not readily available in the corpus and therefore they have to be somehow retrieved and approximated. Given the data sparseness and the parameter estimation difficulties, it is not surprising that far more symbolic than probabilistic solutions have been proposed for the automatic interpretation of compound nouns. With the exception of Wu (1993) and Lauer (1995) who use probabilistic models for compound noun interpretation (see Section 7.6 for details), most algorithms rely on hand crafted knowledge bases or dictionaries which contain detailed semantic information for each noun; a sequence of rules exploit the knowledge base in order to choose the correct interpretation for a given compound (Finin 1980; Leonard 1984; McDonald 1982; Vanderwende 1994).

In what follows we develop a probabilistic model for the interpretation of nominalizations. We focus solely on nominalizations whose prenominal modifier is either the underlying subject or direct object of verb corresponding to the deverbal compound head. In other words, we focus on examples like (7.3a,b) and ignore for the moment nominalizations whose heads correspond to verbs taking prepositional complements (see example (7.3c)). Nominalizations are attractive from an empirical perspective: the amount of relations is small (i.e., subject or object, at least if one focuses on direct objects only) and fairly uncontroversial (see the discussion above). Although the relations are not attested in the corpus, they can be retrieved and approximated through parsing. The probabilistic interpretation of nominalizations can provide a lower bound for the difficulty of the interpretation task: if we cannot interpret nominalizations successfully there is little hope for modeling more complex semantic relations stochastically (see the examples in (7.1)).

The probabilistic model presented in Chapters 4 and 5 will be also used for the interpretation of nominalizations (see Section 7.2). Our approach relies on the simplifying assumption that the relation of the nominalized head and its modifier noun can be approximated by the relation of the latter with the verb from which the head is derived. This approach works insofar as the verb-argument relations from which the nominalizations are derived are attested in the corpus. We show that a large number of verb-argument configurations do not occur in the corpus, something which is not surprising considering the ease with which novel compounds are created (see the discussion in the previous chapter). We estimate the frequencies of unseen

verb-argument pairs by experimenting with three types of smoothing techniques proposed in the literature (back-off smoothing, class-based smoothing, and similarity-based smoothing, see Section 7.2.2) and show that their combination achieves high performance (see Experiments 16 and 17 in Sections 7.3 and 7.4, respectively).

7.2. The Model

Given a nominalization, our goal is to develop a procedure to infer whether the modifier stands in a subject or object relation with respect to the head noun. In other words, we need to assign probabilities to two different relations: SUBJ and OBJ. As in Chapters 4 and 5 we view the choice of a relation *rel* for a compound $n_1 n_2$ as the joint probability $P(n_1, n_2, rel)$. Assuming the ordering $\langle n_1, n_2, rel \rangle$, for each relation *rel* we calculate the simple expression given in (7.6).

$$P(n_1, n_2, rel) = P(n_1) \cdot P(n_2|n_1) \cdot P(rel|n_1, n_2) \tag{7.6}$$

Note that the probabilities $P(n_1)$ and $P(n_2|n_1)$ can be ignored, as they are a constant. The estimation of $P(rel|n_1, n_2)$ is shown in (7.7). Since we have a choice between two outcomes we will use a likelihood ratio to compare the two relation probabilities (Hindle and Rooth 1993; Mosteller and Wallace 1964). In particular, we will compute the log of the ratio of the probability $P(\text{OBJ}|n_1, n_2)$ to the probability $P(\text{SUBJ}|n_1, n_2)$. We will call this log-likelihood ratio the argument relation (*RA*) score (see (7.8)).

$$P(rel|n_1, n_2) = \frac{f(n_2, rel, n_1)}{f(n_1, n_2)} \tag{7.7}$$

$$RA(rel, n_1, n_2) = \log_2 \frac{P(\text{OBJ}|n_1, n_2)}{P(\text{SUBJ}|n_1, n_2)} \tag{7.8}$$

Notice, however, that we cannot read off $f(n_2, rel, n_1)$ directly from the corpus. What we can obtain from a corpus (through parsing) is the number of times a noun is the object or the subject of a given verb. By making the simplifying assumption that the relation of the nominalized head and its modifier noun is the same as the relation between the latter and the verb from which the head is derived, (7.7) is rewritten as follows:

$$P(rel|n_1, n_2) \approx \frac{f(v_{n_2}, rel, n_1)}{\sum_i f(v_{n_2}, rel_i, n_1)} \tag{7.9}$$

Here, $f(v_{n_2}, rel, n_1)$ is the frequency with which the modifier noun n_1 is found in the corpus as the subject or object of v_{n_2}, the verb from which the head noun is derived. The sum $\sum_i f(v_{n_2}, rel_i, n_1)$ is a normalization factor.

7.2.1. Parameter Estimation

7.2.1.1. Verb-argument Tuples

In order to estimate the frequency $f(v_{n_2}, rel, n_1)$ we need to obtain counts of verb-object and verb-subject tuples. Verb-argument relations were also crucial for the estimation of the parameters of the model presented in Chapter 5. Recall from Section 5.2.2.1 that tuples were extracted from a part-of-speech tagged and lemmatized version of the BNC (100 million words) which was automatically parsed by Abney's (1996) chunk parser. The parser's output was post-processed to eliminate erroneous parses by discarding tuples that did not contain verbs and tuples with arguments other than nouns. Furthermore, we discarded tuples containing verbs or nouns with a BNC frequency of one (see Table 5.3 in Section 5.2.2.1). The tuples obtained from this study were also used for the estimation of $P(rel|n_1, n_2)$ (see equation (7.9)).

7.2.1.2. The Data

So far we have been using the term nominalization to refer to two word compounds whose head is derived from a verb. Morphologically speaking, nominalization is a word formation process by which a noun is derived from a verb, usually by means of suffixation (Quirk et al. 1985). A list of deverbal suffixes (i.e., suffixes forming nouns by attaching to verb bases) is given in Table 7.1. Nominalizations can be also created by *conversion*. Conversion is the word formation process whereby an item is adapted or converted to a new word-class without the addition of an affix (Quirk et al. 1985: 1,009). Examples of conversion are shown in Table 7.2.

It is beyond the scope of the present study to develop an algorithm which automatically detects nominalizations in a corpus. In the experiments described in the subsequent sections compounds with deverbal heads were obtained as follows:

1. A dictionary of deverbal nouns was created using two sources: (a) NOMLEX (Macleod et al. 1998), a dictionary of nominalizations containing 827 lexical entries (see Chapter 2 for details on NOMLEX) and (b) CELEX (Burnage 1990), a general morphological dictionary, which contains 5,111 nominalizations (see also Chapter 2 for details); both dictionaries list the verbs from which the nouns are derived. Sample dictionary entries are given in Tables 7.1 and 7.2.

2. Candidate nominalizations were obtained from the compounds acquired from the BNC (see Chapter 6) by selecting noun-noun sequences whose head (i.e., rightmost noun) was one of the deverbal nouns contained either in CELEX or NOMLEX. The procedure resulted in 172,797 potential types of nominalizations.

From these candidate nominalizations a random sample of 1,277 tokens was selected. The sample was manually inspected and compounds with modifiers whose relation to the head

Table 7.1: Deverbal suffixes

Suffix	Nominalization
-ER	drink → drinker
-OR	direct → director
-ANT	disinfect → disinfectant
-EE	employ → employee
-ATION	educate → education
-MENT	arrange → arrangement
-AL	refuse → refusal
-ING	hire → hiring

Table 7.2: Conversion

Verb	→	Noun
release	→	release
arrest	→	arrest
compromise	→	compromise
attempt	→	attempt

noun was other than subject or object were discarded. In particular, nominalizations were discarded if: (a) the relation between the head and the modifier was any of the semantic relations listed in (7.1) (e.g., CAUSE, HAVE, MAKE); these compounds represented 28.0% of the sample or (b) the head was derived from verbs taking prepositional objects (see example (7.3c)); these nominalizations represented 9.20% of the sample. After manual inspection the sample contained 796 nominalizations (62.8% of the initial sample) and 418 distinct deverbal nouns (i.e., compound heads). From these, 596 tokens were used as training data for the experiments described in Sections 7.3 and 7.4. The remaining 200 nominalizations were retained as test data and also in order to evaluate whether human judges can reliably disambiguate the argument relation between the nominalized head and its modifier (see Section 7.2.4).

7.2.1.3. Mapping

In order to estimate the frequency, $f(v_{n_2}, rel, n_1)$, the nominalized heads were mapped to their corresponding verbs. Inspection of the frequencies of the verb-argument tuples contained in our training sample (596 tokens) revealed that 372 verb-noun pairs had a verb-object frequency of zero in the corpus. Similarly, 378 verb-noun pairs had a verb-subject frequency of zero. Furthermore, a total of 287 tuples were not attested at all in the BNC either in a verb-object or verb-subject relation. This finding is perhaps not surprising given the productivity of compounding. Considering the ease with which novel compounds are created, it is to be expected that some verb-argument configurations will not occur in the training corpus.

We recreated the frequencies of unseen verb-argument pairs by experimenting with three types of smoothing techniques proposed in the literature: back-off smoothing (Katz 1987), class-based smoothing (Lauer 1995; Resnik 1993), and similarity-based smoothing (Dagan et al. 1999; Grishman and Sterling 1994). We present these three smoothing variants in Section 7.2.2. In Section 7.2.3 we introduce an algorithm that uses smoothed verb-argument tuples to arrive at the interpretation of nominalizations.

7.2.2. Smoothing

7.2.2.1. Back-off Smoothing

Back-off n-gram models were initially proposed by Katz (1987) for speech recognition but have been also successfully used to disambiguate the attachment site of structurally ambiguous PPs (Collins and Brooks 1995). The main idea behind back-off smoothing is to adjust maximum likelihood estimates like (7.7) so that the total probability of observed word co-occurrences is less than one, leaving some probability mass to be redistributed among unseen co-occurrences. In general, the frequency of observed word sequences is discounted using Good-Turing's estimate (see Katz 1987 and Church and Gale 1991 for details on Good-Turing estimation) and the probability of unseen sequences is estimated using lower order conditional distributions. Assuming that the denominator $f(v_{n_2}, rel, n_1)$ in (7.7) is zero we can approximate $P(rel|n_1, n_2)$ by backing-off to $P(rel|n_1)$:

$$P(rel|n_1, n_2) = \alpha \frac{f(rel, n_1)}{f(n_1)} \tag{7.10}$$

Here, α is a normalization constant which ensures that the probabilities sum to one. If the frequency $f(rel, n_1)$ is also zero, backing-off continues by making use of $P(rel)$.

7.2.2.2. Class-based Smoothing

An instantiation of class-based smoothing was already presented in Section 6.9.1. Recall that we captured how likely it is for two nouns to form a compound by substituting the head and modifier by the concepts with which they are represented in a taxonomy. The frequency of the concept pair was recreated by taking into account compounds with same concepts in the taxonomy. Generally speaking, class-based smoothing recreates co-occurrence frequencies based on information provided by lexical resources such as WordNet (Miller et al. 1990) or Roget's thesaurus. In the case of verb-argument tuples we use taxonomic information to estimate the frequencies $f(v_{n_2}, rel, n_1)$ by substituting the word n_1 occurring in an argument position by the concept with which it is represented in the taxonomy (Resnik 1993). So, $f(v_{n_2}, rel, n_1)$ can be estimated by counting the number of times the concept corresponding to n_1 was observed as the argument of the verb v_{n_2} in the corpus.

Recall from Section 6.9.1 that words in a taxonomy typically belong to more than one conceptual class. Because an argument of a verb can generally be the realization of one of several conceptual classes, counts of verb-argument configurations are constructed for each conceptual class by dividing the contribution from the argument by the number of classes it

Table 7.3: Frequency estimation for *group registration* using WordNet

Verb	Class	$f(v_{n_2}, \text{OBJ}, n_1)$	$f(v_{n_2}, \text{SUBJ}, n_1)$
register	⟨abstraction⟩	16.26	7.28
register	⟨entity⟩	14.10	4.50
register	⟨object⟩	8.02	1.56
register	⟨set⟩	.65	.07
register	⟨substance⟩	.70	.08

belongs to (Lauer 1995; Resnik 1993):

$$f(v_{n_2}, rel, c) \approx \sum_{n_1' \in c} \frac{f(v_{n_2}, rel, n_1')}{|classes(n_1')|} \tag{7.11}$$

Here, $f(v_{n_2}, rel, n_1')$ is the number of times the verb v_{n_2} was observed with concept $c \in classes(n_1')$ bearing the argument relation *rel* (i.e., subject or object) and $|classes(n_1')|$ is the number of conceptual classes n_1' belongs to.

Consider for example the tuple *register group* (derived from the compound *group registration*) which is not attested in the BNC. The word *group* has two senses in WordNet and belongs to five conceptual classes (⟨abstraction⟩, ⟨entity⟩, ⟨object⟩, ⟨set⟩, and ⟨substance⟩). This means that the frequency $f(v_{n_2}, rel, c)$ will be constructed for each of the five classes, as shown in Table 7.3. Suppose for example that we see the tuple *register patient* in the corpus. The word *patient* has two senses in WordNet and belongs to seven conceptual classes (⟨case⟩, ⟨person⟩, ⟨life form⟩, ⟨entity⟩, ⟨causal agent⟩, ⟨sick person⟩, ⟨unfortunate⟩) one of which is ⟨entity⟩. This means that we will increment the observed co-occurrence count of *register* and ⟨entity⟩ by $\frac{1}{7}$. Since we do not know which is the actual class of the noun *group* in the corpus we weight the contribution of each class by taking the average of the constructed frequencies for all five classes:

$$f(v_{n_2}, rel, n_1) = \frac{\sum_{c \in classes(n_1)} \sum_{n_1' \in c} \frac{f(v_{n_2}, rel, n_1')}{|classes(n_1')|}}{|classes(n_1)|} \tag{7.12}$$

Following (7.12) the frequencies $f(register, \text{OBJ}, group)$ and $f(register, \text{SUBJ}, group)$ are $\frac{39.73}{5}$ and $\frac{13.49}{5}$, respectively. Note that the estimation of the frequency $f(v_{n_2}, rel, n_1)$ (see equations (7.11) and (7.12)) crucially relies on the simplifying assumption that the argument of a verb is distributed evenly across its conceptual classes.

7.2.2.3. Similarity-based Smoothing

Similarity-based smoothing relies on the assumption that if a word w_1' is "similar" to word w_1, then w_1' can provide information about the frequency of unseen word pairs involving w_1 (Dagan et al. 1999). A key feature of similarity-based smoothing is the function which measures distributional similarity from co-occurrence frequencies. Several measures of word similarity which can be derived from word co-occurrences have been proposed in the literature, (see Dagan et al. 1999 and Lee 1999 for overviews), providing an alternative to taxonomies such as WordNet.

We have experimented with two measures of distributional similarity derived from co-occurrence frequencies: the Jensen-Shannon divergence and the confusion probability. The choice of these two measures was motivated by work described in Dagan et al. (1999) where the Jensen-Shannon divergence outperformed related similarity measures (such as the confusion probability or the L norm) on a word sense disambiguation task which used verb-object pairs. The confusion probability has been used by several authors in order to smooth word co-occurrence probabilities (Essen and Steinbiss 1992; Grishman and Sterling 1994). Grishman and Sterling (1994) in particular employed the confusion probability to recreate the frequencies of verb-noun co-occurrences where the noun was the object or the subject of the verb in question. In the following we describe these two similarity measures and show how they can be used to recreate the frequencies for unseen verb-argument tuples (for a more detailed description see Dagan et al. 1999).

Confusion Probability. The confusion probability P_C is an estimate of the probability that word w_1' can be substituted by word w_1, in the sense of being found in the same contexts. In other words, the metric indicates how probable it is for word w_1' to occur in contexts in which word w_1 occurs. A large confusion probability value indicates that the two words w_1' and w_1 appear in similar contexts.

$$P_C(w_1|w_1') = \sum_s P(w_1|s)P(s|w_1') \tag{7.13}$$

Here, $P_C(w_1'|w_1)$ is the probability that word w_1' occurs in the same contexts s as word w_1, averaged over these contexts. Given a tuple of the form w_1, rel, w_2 we can either treat w_1, rel as context and smooth over the noun w_2 or rel, w_2 as context and smooth over the verb w_1. We opted for the latter for theoretic reasons since it is the verb which imposes the semantic restrictions on its arguments and not vice versa. The idea that semantically similar verbs have similar subcategorizational and selectional patterns is by no means new, and has been extensively argued for by Levin (1993). By taking verb-argument tuples into consideration (7.13) is

rewritten as follows:

$$(7.14) \quad P_C(w_1|w_1') = \sum_{rel,w_2} P(w_1|rel,w_2)P(rel,w_2|w_1') = \sum_{rel,w_2} \frac{f(w_1,rel,w_2)}{f(rel,w_2)} \frac{f(w_1',rel,w_2)}{f(w_1')}$$

The confusion probability can be computed efficiently since it involves summation only over the common contexts rel, w_2.

Jensen-Shannon Divergence. The Jensen-Shannon divergence J is an information-theoretic measure. It recasts the concept of distributional similarity into a measure of the "distance" (i.e., dissimilarity) between two probability distributions. The value of the Jensen-Shannon divergence ranges from zero for identical distributions to $\log 2$ for maximally different distributions.

$$(7.15) \quad J(w_1,w_1') = \frac{1}{2}\left[D\left(w_1 \left\| \frac{w_1+w_1'}{2}\right.\right) + D\left(w_1' \left\| \frac{w_1+w_1'}{2}\right.\right)\right]$$

$$(7.16) \quad D(w_1\|w_1') = \sum_{rel,w_2} P(rel,w_2|w_1) \log \frac{P(rel,w_2|w_1)}{P(rel,w_2|w_1')}$$

Here, D in (7.15) is the Kullback-Leibler divergence, a measure of the dissimilarity between two probability distributions (see equation (7.16)) and $(w_1+w_1')/2$ is a shorthand for the average distribution:

$$(7.17) \quad \frac{1}{2}(P(rel,w_2|w_1) + P(rel,w_2|w_1'))$$

Similarly to the confusion probability, the computation of J depends only on the common contexts rel, w_2. Recall that the Jensen-Shannon divergence is a dissimilarity measure. The dissimilarity measure is transformed to a similarity measure using a weight function $W_J(w,w_1')$:

$$(7.18) \quad W_J(w_1,w_1') = 10^{-\beta J(w_1,w_1')}$$

The parameter β controls the relative influence of the neighbors (i.e., distributionally similar words) closest to w_1: if β is high, only neighbors extremely close to w_1 contribute to the estimate, whereas if β is low distant neighbors also contribute to the estimate.

We estimate the frequency of an unseen verb-argument tuple by taking into account the similar w_1s and the contexts in which they occur (Grishman and Sterling 1994):

$$(7.19) \quad f_s(w_1,rel,w_2) = \sum_{w_1'} \mathrm{sim}(w_1,w_1') f(w_1',rel,w_2)$$

Given a set of nominalizations n_1 n_2:

1. map the head noun n_2 to the verb v_{n_2} from which it is derived;
2. retrieve frequencies $f(verb_{n_2}, \text{OBJ}, n_1)$ and $f(verb_{n_2}, \text{SUBJ}, n_1)$ from the BNC;
3. **if** $f(verb_{n_2}, \text{OBJ}, n_1) < k$ **then**
   recreate $f_s(verb_{n_2}, \text{OBJ}, n_1)$;
4. **if** $f(verb_{n_2}, \text{SUBJ}, n_1) < k$ **then**
   recreate $f_s(verb_{n_2}, \text{SUBJ}, n_1)$;
5. calculate probabilities $P(\text{OBJ}|n_1, n_2)$ and $P(\text{SUBJ}|n_1, n_2)$;
6. compute $RA(rel, n_1, n_2)$;
7. **if** $RA \leq j$ **then**
   n_1 is the subject of n_2;
8. **else**
   n_1 is the object of n_2.

Figure 7.1: Disambiguation algorithm for nominalizations

Here, $sim(w_1, w_1')$ is a function of the similarity between w_1 and w_1'. In our experiments $sim(w_1, w_1')$ was substituted by the confusion probability $P_C(w_1|w_1')$ and the Jensen-Shannon divergence $W_J(w_1, w_1')$.

7.2.3. The Algorithm

The disambiguation algorithm for nominalizations is summarized in Figure 7.1. The algorithm uses verb-argument tuples in order to infer the relation holding between a modifier and its nominalized head. When the co-occurrence frequency of the verb-argument relation is zero, verb-argument tuples are smoothed using one of the methods described in Section 7.2.2.

Once frequencies for verb-argument relations (either actual or reconstructed through smoothing) have been obtained, the argument relation (*RA*) score determines the relation between the head n_2 and its modifier n_1 (see Section 7.2). The sign of the *RA* score indicates which relation, subject or object, is more likely: a positive *RA* score indicates an object relation, whereas a negative score indicates a subject relation. Depending on the task and the data at hand we can require that an object or subject analysis is preferred only if *RA* exceeds a certain threshold j (see steps 7 and 8 in Figure 7.1). We can also impose a threshold k on the type of verb-argument tuples we smooth. If for instance we know that the parser's output is noisy, then we might choose to smooth not only unseen verb-argument pairs but also pairs with non-zero

Table 7.4: *RA* scores for verb-argument tuples extracted from the BNC

Verb-noun	$f(v_{n2}, \text{OBJ}, n_1)$	$f(v_{n2}, \text{SUBJ}, n_1)$	*RA*
administer student	0	0	.96
establish unit	22	1	.55
promote government	3	10	−1.73

corpus frequencies (e.g., $f(verb_{n_2}, rel, n_1) \geq 1$, see steps 3 and 4 in Figure 7.1).

As an example consider the compound *student administration*: its corresponding verb-noun configuration (i.e., *administer student*) is not attested in the BNC. This is a case where we need smoothed estimates for both $f(v_{n2}, \text{OBJ}, n_1)$ and $f(v_{n2}, \text{SUBJ}, n_1)$. The recreated frequencies using the class-based smoothing method described in Section 7.2.2.2 are 5.06 and 2.59, respectively, yielding an *RA* score of .96 which means that it is more likely that *student* is the object of *administration* (see Table 7.4). Consider now the compound *unit establishment*: here, we have very little evidence in the corpus with respect to the verb-subject relation (see Table 7.4, where $f(establish, \text{SUBJ}, unit) = 1$). Assuming we have set the threshold k to 2 (see steps 4 and 5 in Figure 7.1) we need only recreate the frequency for the subject relation (e.g., 14.99 using class-based smoothing). The resulting *RA* score is again positive (see Table 7.4) which indicates that there is a greater probability for *unit* to be the object of *establishment* than for it to be the subject. Finally, consider the compound *government promotion*: both subject and object relations are represented in the BNC (see Table 7.4) in which case no smoothing is involved; we need only calculate the *RA* score (see step 6 in Figure 7.1) which is negative, indicating that *government* is more likely to be the subject of *promotion* than its object.

7.2.4. Agreement

Before attempting to interpret nominalizations automatically we evaluated if humans can reliably decide whether a nominalization receives a subject or object interpretation. Two judges were presented with 200 nominalizations and were asked to decide whether the modifier is the subject or object of a given nominalized head. The judges were given a page of guidelines (see Section A.3 in Appendix A) but no prior training. The nominalizations were disambiguated in context: the judges were given the corpus sentence in which the nominalization occurred together with the previous and following sentence. The judges' agreement was measured using the Kappa coefficient (Cohen 1960, see Chapter 2 for details).

The judges' agreement on the disambiguation task was $K = .78$ ($N = 200$, $k = 2$). The agreement was good given that the judges were provided with minimal instructions and no prior training. However, note that despite the fact that context was provided to aid the disambiguation task the judges were not in complete agreement. Recall from Section 2.5.1 that $K = 1$ if there is complete agreement. This points to the intrinsic difficulty of the task at hand.

Argument relations and consequently selectional restrictions are influenced by several pragmatic factors which may not be readily inferred from the immediate context (see the discussion in Section 7.5). In the following, we propose a method which raises a greater challenge: the interpretation of nominalizations without taking context into account.

7.3. Experiment 16: Smoothing Variants

7.3.1. Method

In Experiment 16 we explore the influence of the different smoothing methods (back-off smoothing, class-based smoothing, and similarity-based smoothing) on the interpretation task. As far as class-based smoothing is concerned we experimented with two concept hierarchies, Roget's thesaurus and WordNet. With respect to the similarity-based smoothing, we examined: (a) the performance of the two different similarity functions (confusion probability and Jensen-Shannon divergence), (b) the size of the vocabulary (e.g., number of verbs used to find the nearest neighbors) and its impact on disambiguation performance, and (c) the effect of the β parameter.

We also investigated whether knowledge of the semantics of the suffix of the nominalized head can improve performance. We ran two versions of the algorithm detailed in Section 7.2.3: in one version the algorithm assumes no prior knowledge about the semantics of the nominalization suffix (see Figure 7.1); in the other version the algorithm estimates the probabilities $P(\text{OBJ}|n_1,n_2)$ and $P(\text{SUBJ}|n_1,n_2)$ only for compounds with nominalization suffixes other than *-er*, *-or*, *-ant*, or *-ee*. For compounds with suffixes *-er*, *-or*, and *-ant* (e.g., *datum holder*, *car collector*, *water disinfectant*) the algorithm defaults to an object interpretation, and for the suffix *-ee* (e.g., *university employee*) the algorithm defaults to a subject analysis. Compounds with heads ending in these four suffixes represented 13.6% of the compounds in the training set and 10.8% of the compounds in the test set.

The parameters of the algorithm were tuned on 596 nominalizations and tested on 200. The 596 nominalizations were also used as training data for finding the optimal parameters for the two parameterized similarity-based smoothing approaches (see Section 7.2.2.3). The algorithm's output was compared to the manual classification (see Section 7.2.4) and precision was computed accordingly. For 61.5% of the nominalizations contained in the test data the modifier was the object of the deverbal head, whereas in the remaining 38.5% the modifier was the subject. This means that a simple heuristic which defaults to an object relation yields a precision of 61.5%. Our algorithm defaults to an object relation when there is no evidence to support either analysis (e.g., when $f(v_{n_2},\text{OBJ},n_1) = f(v_{n_2},\text{SUBJ},n_1)$, see Step 8 in Figure 7.1).

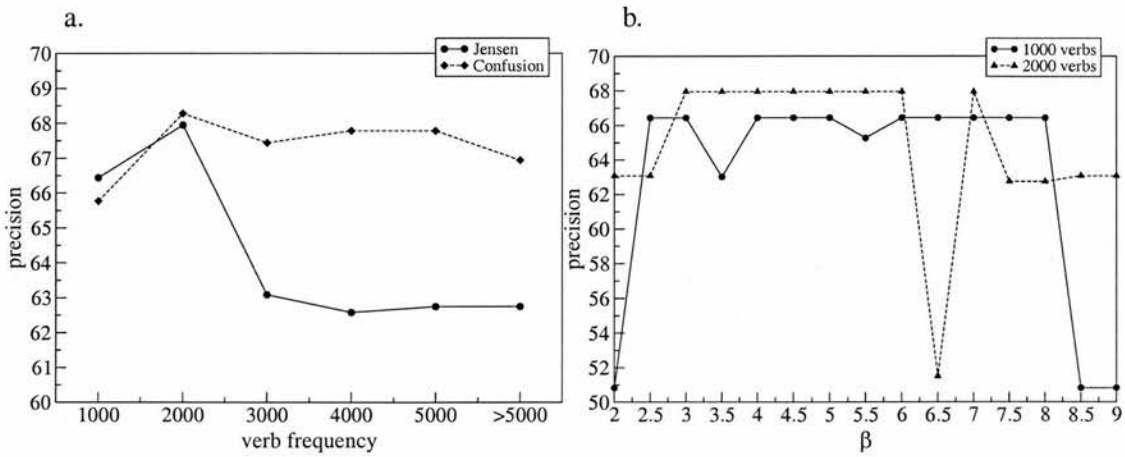

Figure 7.2: Parameter settings for P_C and J divergence

Table 7.5: 10 closest words to verb *accept* for P_C

Verbs					
1,000	2,000	3,000	4,000	5,000	> 5,000
accept	decline	decline	decline	decline	incl
refuse	accept	tender	tender	re-issued	decline
reject	refuse	accept	abdicate	co-manage	re-issued
submit	delegate	table	accept	tender	co-manage
endorse	reject	disclaim	table	oversubscribe	tender
approve	repudiate	plate	wangle	backdate	goodwill
issue	hitch	shirk	disclaim	abdicate	oversubscribe
implement	shoulder	refuse	plate	accept	pre-arrange
acknowledge	delegate	proffer	shirk	table	backdate
incur	ratify	apportion	disdain	wangle	abdicate

7.3.2. Results

Before reporting the results of the disambiguation task, we describe our initial experiments on finding the optimal parameter settings for the two similarity-based smoothing methods.

Figure 7.2a shows how performance on the disambiguation task varies with respect to the number and frequency of verbs over which the similarity function is calculated. The y axis in Figure 7.2a shows how performance on the training set varies (for both P_C and J divergence) when verb-argument pairs are selected for the 1,000 most frequent verbs in the corpus, the 2,000 most frequent verbs in the corpus, etc. (x axis). The best performance for both similarity functions is achieved with the 2,000 most frequent verbs. Furthermore, performance between J and P_C is comparable (68.0% and 68.3%, respectively). Another important observation is that performance deteriorates less severely for P_C than for J as the number of verbs increases: when all verbs for which verb-argument tuples are extracted from the BNC are used precision for P_C is 66.9%, whereas precision for J is 62.8%. These results are perhaps unsurprising:

Table 7.6: Disambiguation performance without nominalization suffixes

Methods	Train (%)	Test (%)
D	59.0 ± 2.01	61.5 ± 3.50
B	63.1 ± 1.98	69.6 ± 3.31
P_C	68.3 ± 1.90	75.8 ± 3.08
J	68.0 ± 1.91	69.1 ± 3.33
Wn	68.0 ± 1.91	72.7 ± 3.20
Ro	65.0 ± 1.95	68.6 ± 3.34

Table 7.7: Disambiguation performance with nominalization suffixes

Methods	Train (%)	Test (%)
D	59.0 ± 2.01	61.5 ± 3.50
B	67.5 ± 1.92	69.6 ± 3.31
P_C	70.6 ± 1.87	76.3 ± 3.06
J	69.0 ± 1.89	69.6 ± 3.31
Wn	70.5 ± 1.87	74.2 ± 3.15
Ro	67.5 ± 1.92	69.6 ± 3.31

Table 7.8: Disambiguation performance on nominalizations with suffixes *-er, -or, -ant, -ee* only

Methods	Train (%)	Test (%)
B	79.0	95.2
P_C	74.1	90.5
J	84.0	95.2
Wn	72.3	81.0
Ro	73.4	81.0

verb-argument pairs with low-frequency verbs introduce noise due to the errors inherent in the partial parser. Table 7.5 shows the 10 closest words to the verb *accept* according to P_C as the number of verbs is varied: the quality of the closest neighbors deteriorates with the inclusion of less frequent verbs.

Finally, we analyzed the role of the parameter β. Recall that β appears in the weight function for the Jensen-Shannon divergence and controls the influence of the most similar words: the contribution of the closest neighbors increases with a high value for β. Figure 7.2b shows how the value of β affects performance on the disambiguation task when the similarity function is computed for the 1,000 and 2,000 most frequent verbs in the corpus. It is clear that performance is low with high or very low β values (e.g., $\beta \in \{2,9\}$). We chose to set the parameter β to five and the results shown in Figure 7.2a have been produced for this value for all verb frequency classes.

Table 7.6 shows how the three types of smoothing, back-off (B), class-based (using WordNet (Wn) and Roget (Ro)), and similarity-based (using confusion probability (P_C) and the Jensen-Shannon divergence (J)), influence performance in predicting the relation between a modifier and its nominalized head. For comparison we also show the performance of the simple strategy of always defaulting to an object relation (D). For the similarity-based methods we report the results obtained with the optimal parameter settings ($\beta = 5$; 2,000 most frequent verbs). The results in Table 7.6 were obtained without taking the semantics of the nominalization suffix (*-er, -or, -ant, -ee*) into account (see Section 7.3.1).

Let us concentrate on the training set first. The back-off method is outperformed by all other methods, although its performance is comparable to class-based smoothing using Roget's

thesaurus (63.1% and 65.0%, respectively). Similarity-based methods outperform concept-based methods, although not significantly (accuracy on the training set was 68.3% for P_C and 68.0% for class-based smoothing using WordNet). Furthermore, the particular concept hierarchy used for class-based smoothing seems to have an effect on disambiguation performance: an increase of 2.9% is obtained by using WordNet instead of Roget's thesaurus. One explanation might be that Roget's thesaurus is too coarse-grained a taxonomy for the task at hand (see Section 6.9.4 for a similar observation). We used the χ^2 statistic to examine whether the observed performance is better than the simple strategy of always choosing an object relation which yields an accuracy of 59.0% in the training data (see *D* in Table 7.6). The proportion of nominalizations classified correctly was significantly greater than 59.0% ($p < .01$) for all methods but back-off (*B*) and Roget (*Ro*).

Similar results were observed on the test set. Again P_C outperforms all other methods achieving a precision of 75.8% (see Table 7.6). The portion of nominalizations classified correctly by P_C was significantly greater than 61.5% ($p < .01$) which was the percentage of object relations in the test set. The second best method is class-based smoothing using WordNet (see Table 7.6). WordNet's performance is also significantly better than the baseline ($p < .05$). The back-off method, class-based smoothing using Roget's thesaurus, and *J* yield comparable results (see Table 7.6).

Table 7.7 shows how each method performs when knowledge about the semantics of the nominalization suffix is taken into account. Recall from Section 7.3.1 that compounds with heads ending in agentive or passive suffixes represented 13.6% of the compounds contained in the training set and 10.8% of the compounds in the test set. A general observation is that knowledge of the semantics of the nominalization suffix does not dramatically influence accuracy. Performance on the test data increases by 1.54% for WordNet, 1.02% for Roget .51%, for *J*, and .62% for P_C. We observe no increase in performance for back-off smoothing (compare Tables 7.6 and 7.7). These results suggest that the semantics of the four suffixes *-er, -or, -ant,* and *-ee* can be successfully retrieved from the corpus. Table 7.8 shows the precision of the five smoothing variants on predicting the argument relation only for nominalizations whose heads have these four suffixes (13.6% of the training data and 10.8% of the test data).

An interesting question is the extent to which any of the different methods agree in their assignments of subject and object relations. We investigated this by calculating the methods' agreement on the training set using the Kappa coefficient (see Section 2.5.1 for details). We calculated the Kappa coefficient for all six pairwise combinations of the five smoothing variants. The results are reported in Table 7.9. The highest agreement is observed for P_C and the class-based smoothing using the WordNet taxonomy ($K = .75$). Agreement between *J* and P_C as well as agreement between WordNet and Roget's thesaurus was rather low ($K = .53$ and $K = .46$, respectively). Note that generally low agreement is observed when back-off is paired with either *J*, P_C, WordNet, or Roget. This is not entirely unexpected given the assumptions

Table 7.9: Agreement between smoothing methods

	B	*J*	P_C	*Wn*
J	.31			
P_C	.26	.53		
Wn	.27	.37	.75	
Ro	.25	.26	.49	.46

Table 7.10: Performance at predicting argument relations

	Train (%)		Test (%)	
Methods	SUBJ	OBJ	SUBJ	OBJ
B	41.6	78.1	38.0	87.8
P_C	47.4	82.9	54.9	87.8
J	34.7	91.2	35.2	88.6
Wn	47.8	82.1	49.3	86.2
Ro	50.6	74.4	46.5	81.3

underlying the different smoothing techniques. Both class-based and similarity-based methods recreate the frequency of unseen word combinations by relying on corpus evidence for words that are distributionally similar to the words of interest. In similarity-based smoothing word similarity is estimated from lexical co-occurrence information, whereas in class-based smoothing similarity is provided by conceptual taxonomies. Back-off smoothing, however, incorporates no notion of similarity: unseen sequences are not estimated using similar conditional distributions, but lower order ones. This also relates to the fact that back-off's performance is lower than WordNet and P_C (see Table 7.6) which suggests that smoothing methods incorporating semantic hypotheses (i.e., the notion of similarity) perform better than methods relying simply on co-occurrence distributions. The agreement values in Table 7.9 further suggest that methods inducing similarity relationships from corpus co-occurrence statistics are not necessarily incompatible with methods which quantify similarity using manually crafted taxonomies and that different smoothing techniques may be appropriate for different tasks.

Table 7.10 shows how the different methods compare for the task of predicting the individual argument relations for the training and test sets without taking suffix semantics into account (see the version of the algorithm in Figure 7.1). A general observation is that all methods are fairly good at predicting object relations. Predicting subject relations is considerably harder: no method exceeds an accuracy of 54.9% (see Table 7.10). One explanation for this is that selectional constraints imposed on subjects can be more easily overridden by pragmatic and contextual factors than those imposed on objects. Another explanation (somewhat supported by our experiments on adjective-noun combinations) is that the parser is better at recognizing verb-object than verb-subject relations. *J* is particularly good at predicting object relations, whereas P_C and class-based smoothing using WordNet seem to yield comparable performances when it comes to predicting subject relations (see Table 7.10).

7.4. Experiment 17: Decision Tree Learning

7.4.1. Method

In Experiment 17 we explore how the combination of the different smoothing methods influences the interpretation accuracy. An obvious question is whether the precision can be increased when combining the five smoothing variants given that they seem to provide complementary information for predicting argument relations. For example, WordNet, Roget's thesaurus, and P_C are relatively good for the prediction of subject relations, whereas the Jensen divergence is best for the prediction of object relations (see Table 7.10). We combined the five information sources using Ripper (Cohen 1996). The decision tree was trained on the 596 nominalizations on which the smoothing methods were trained and tested on the 200 unseen nominalizations for which the inter-judge agreement was previously calculated (see Section 7.2.4).

7.4.2. Results

Table 7.11 shows how the precision on the interpretation of nominalizations varies when pairs of the five smoothing variants are used. The decision tree attains a precision of 79.9% on the test set when J is paired with back-off, P_C, Roget, or WordNet (see Table 7.11), achieving thus an increase of 4.10% over 75.8%, P_C's performance on the test data (see Table 7.6). Combination of back-off with P_C or WordNet yields an increase over back-off's individual performance (2.60% and 3.60% when back-off is paired with P_C and WordNet, respectively), without, however, outperforming P_C's individual accuracy on the test data (see Tables 7.11 and 7.6). Pairing back-off with Roget performs .04% better than the baseline of 61.5% (see Table 7.11). Combination of WordNet with Roget yields an increase of .50% over WordNet and 4.60% over Roget alone (see Table 7.11). Finally, combination of WordNet with P_C yields an accuracy of 73.7% which outperforms WordNet alone by 1.00%, but is in fact 2.10% lower than P_C's individual accuracy. This is not surprising since P_C and WordNet tend to agree in their assignments of subject and object relations (see the methods' agreement in Table 7.9) and therefore their combination is not expected to be very informative.

Combining a class-based method with a similarity-based one performs as well as combining two similarity-based methods (see J, Wn, and P_C, J in Table 7.11). This means that we could interpret nominalizations using information inherent in the corpus without making external assumptions with regard to how concepts and their similarity are represented. A similar result is observed when combinations of triples of smoothing variants are taken into account. Combination of the three non-taxonomic smoothing variants yields the same performance (see methods B, J and P_C in Table 7.12) as combination of taxonomic smoothing variants with non-taxonomic ones (see for example Ro, B, J or Wn, Ro, J in Table 7.12). Combination of any two methods with J achieves an accuracy of 79.9%. This is not surprising, since J alone performs best at predicting object relations (see Table 7.10) which outnumber subject relations both in

Table 7.11: Decision trees with two smoothing variants

Methods	Train (%)	Test (%)
B, *J*	78.5 ± 1.68	79.9 ± 2.88
B, P_C	68.5 ± 1.91	72.2 ± 3.23
P_C, *J*	78.5 ± 1.68	79.9 ± 2.88
B, *Ro*	64.6 ± 1.96	61.9 ± 3.50
J, *Ro*	78.5 ± 1.68	79.9 ± 2.88
P_C, *Ro*	68.5 ± 1.91	72.2 ± 3.23
B, *Wn*	68.8 ± 1.90	73.2 ± 3.19
J, *Wn*	78.5 ± 1.68	79.9 ± 2.88
P_C, *Wn*	69.8 ± 1.88	73.7 ± 3.17
Wn, *Ro*	68.8 ± 1.90	73.2 ± 3.19

Table 7.12: Decision trees with three smoothing variants

Methods	Train (%)	Test (%)
B, *J*, P_C	78.5 ± 1.68	79.9 ± 2.88
Ro, *B*, *J*	78.5 ± 1.68	79.9 ± 2.88
Ro, *B*, P_C	68.5 ± 1.91	72.2 ± 3.23
Ro, *J*, P_C	78.5 ± 1.68	79.9 ± 2.88
Wn, *B*, *J*	78.5 ± 1.68	79.9 ± 2.88
Wn, *B*, P_C	69.8 ± 1.88	73.7 ± 3.17
Wn, *J*, P_C	78.5 ± 1.68	79.9 ± 2.88
Ro, *Wn*, *B*	68.8 ± 1.90	73.2 ± 3.19
Ro, *Wn*, *J*	78.5 ± 1.68	79.9 ± 2.88
Ro, *Wn*, P_C	69.8 ± 1.88	73.7 ± 3.17

Table 7.13: Decision trees with four and five smoothing variants

Methods	Train (%)	Test (%)
B, *J*, P_C, *Ro*	78.5 ± 1.68	79.9 ± 2.88
B, *J*, P_C, *Wn*	79.7 ± 1.65	79.9 ± 2.88
Ro, *Wn*, *B*, *J*	78.5 ± 1.68	79.9 ± 2.88
Ro, *Wn*, *B*, P_C	68.8 ± 1.90	73.7 ± 3.17
Ro, *Wn*, *J*, P_C	79.7 ± 1.65	79.9 ± 2.88
B, *J*, P_C, *Ro*, *Wn*	79.7 ± 1.65	79.9 ± 2.88

our training and test data.

Accuracy on the interpretation task remains steadily good when quadruples of the five smoothing variants are used (i.e., 79.9%, see Table 7.13). Note that the absence of *J* from the combination of smoothing variants, results in a decrease of 6.70% over the best performance of 79.9% (see *Ro*, *Wn*, *B*, P_C in Table 7.13). This is not unexpected given that *J* is the best predictor for object relations and P_C and WordNet behave similarly with respect to their interpretation decisions. An accuracy of 79.9% is also attained when all five smoothing methods are combined (see Table 7.13).

We further analyzed the decision tree learner's performance at predicting object and subject relations. This information is displayed in Table 7.14, where we show how performance varies when the best two, three, four, and five methods are combined. Since the highest performance (i.e., 79.9%) was achieved by several method combinations, we display one combination for each type of decision tree (e.g., decision tree with two, three or four methods, etc.). Precision at predicting subject relations increases when smoothing variants are combined (compare Tables 7.14 and 7.10). In fact, an increase of 8.40% over 62.9% (the subject accuracy resulting from the combination of two or three methods) is obtained for subject relations when more than three methods are combined, although accuracy at predicting object relations still remains

Table 7.14: Ripper's performance at predicting argument relations

	Train (%)		Test (%)	
Methods	SUBJ	OBJ	SUBJ	OBJ
B, J	62.9	89.5	64.8	88.6
Ro, B, J	62.9	89.5	64.8	88.6
J, P_C, Ro, Wn	71.4	84.5	73.2	83.7
B, J, P_C, Ro, Wn	71.4	84.5	73.2	83.7

higher (see Table 7.14).

The results reported in Tables 7.11–7.13 were obtained without having access to the semantic information of the nominalization suffix. As explained in Section 7.3.2 a slight increase in performance is obtained when suffix semantics are taken into account. When all five smoothing variants are combined together with knowledge about the semantics of the nominalization suffix, a small increase of.50% is achieved. (80.2% $\pm$ 1.63 on training data and 80.4% on test data for B, J, P_C, Ro, Wn).

In sum, our results show that combination of the different smoothing variants (using decision tree learning) achieves better results than each individual method. Good performance (i.e., 79.9%) can be achieved by taking only similarity-based methods into account (see P_C, J in Table 7.11). However, the best performance (with regard to predicting object and subject relations) is achieved when similarity-based methods are combined with class-based methods. For example, the combination of J with P_C, Ro, and Wn performs best at predicting subject relations achieving a significant increase of 35.2% ($p < .01$) over back-off's individual performance of 38.0%, an increase of 18.3% ($p < .01$) over P_C's accuracy of 54.9%, an increase of 38.0% ($p < .01$) over J's accuracy of 35.2%, an increase of 23.9% ($p < .01$) over WordNet, and an increase of 26.7% ($p < .01$) over Roget (compare Tables 7.14 and 7.10).

7.5. Discussion

We have described a statistical model which interprets nominalizations by taking into account the occurrence of verb-argument tuples in the corpus. We showed that a simple algorithm which combines information about the distributional properties of words and their syntactic relations and domain independent symbolic knowledge (e.g., WordNet) achieves a performance of approximately 80% on the interpretation task which is significantly higher than the baseline of 61.5%. This is an important result considering the simplifications in the system and the sparse data problems encountered in the estimation of the probability $P(rel|n_1,n_2)$. In this chapter we provided further evidence for the surface cueing approach. We have shown that surface grammatical relations acquired from the corpus can be used to approximate the relation of a deverbal compound head to its modifier and furthermore that when grammatical relations are unattested in the corpus, they can be successfully recreated using smoothing techniques that

rely on linguistic hypotheses (i.e., words display meaning similarities when found in similar contexts).

We have further demonstrated how the model presented in Chapters 4 and 5 performs under conditions of data sparseness. We compensated for the lack of insufficient distributional information by taking advantage of smoothing methods that "recreate" the frequencies of word combinations either by simply relying on verb argument tuples extracted from a large corpus (e.g., back-off, similarity-based smoothing) or in conjunction with taxonomic information (e.g., WordNet). As in Chapter 6, we explored the contribution of several information sources (e.g., context, taxonomies, parsing) using a uniform methodology, i.e., decision tree learning which seems appropriate both for the acquisition and interpretation of compounds.

The interpretation of compound nouns is important for several NLP tasks, notably machine translation. Consider the compound *satellite observation* (taken from (7.4a)) which may mean "observation by satellite" or "observation of satellites". In order to translate *satellite observation* into Spanish, we have to work out whether *satellite* is the subject or object of the verb *observe*. In the first case *satellite observation* translates as *observación por satelite* (observation by satellite), whereas in the latter it translates as *observación de satelites* (observation of satellites).

To a certain extent the difficulty of interpreting nominalizations is due to their context dependence. The algorithm presented in the previous sections does not take into account the discourse context in which the nominalizations occur. Consider for instance the compound *film interpretation*. One might argue that the preferred interpretation is "a film which someone interprets". However, the preference for an object interpretation can be easily overridden in context as shown in (7.20a): here it is the film which is doing the interpreting. Note that this is a case where the selectional restrictions of the verb *interpret* underlying the noun *interpretation* are violated: the object (i.e., the film) is doing the interpreting instead of its creator (i.e., the film director). Future work needs to determine the type of discourse clues which are important for the disambiguation of nominalizations. For example, the fact that *interpretation* in (7.20a) is modified by an *of*-PP strongly indicates that *book*, instead of *film*, is its subject.

(7.20) a. The following videos will be screened: BLACK SKIN, WHITE MASKS Part 2. A [film_N interpretation_N] of the book which satirizes black assimilation into white society. (Director: Calvin Brown)

b. Of course, none of this means that the equipment is taking anything away from the chef's own individual skills which are irreplaceable. What it does ensure is that the chef has complete control over some of the most vital tools of his trade, with [computer_N guidance_N] as an important aid.

In sentence (7.20b) the compound *computer guidance* receives a subject interpretation (e.g., the computer guides the chef). The algorithm cannot detect that the *computer* here is ascribed animate qualities and opts for the most likely interpretation (i.e., an object analysis).

In some cases the modifier stands in a metonymic relation to its head. Consider the examples in sentence (7.21) where the nominalizations *industry reception* and *market acceptance* can be thought of as instances of the metonymic schema "Whole for Part" (Lakoff and Johnson 1980). In example (7.21a) it is the industry as a whole which receives the guests instead of LASMO which is one of its parts, and in (7.21b) the modifier *market* in *market acceptance* refers to the opinion leaders who are part of the market.

(7.21) a. The final evening saw more than 300 guests attend an [industry$_N$ reception$_N$], hosted by LASMO.

b. Marketers interested in the development and introduction of new products will be particularly interested in the attitude of opinion leaders to these products, for their general [market$_N$ acceptance$_N$] can be slowed down or speeded up by the views of such people.

The observation that discourse context may favor infrequent interpretations is by no means new or particular to nominalizations. Copestake and Lascarides (1997) and Lascarides and Copestake (1998) make the same observation for a variety of constructions such as compound nouns, adjective-noun combinations, and verb-argument relations. Consider the sentences in (7.22)–(7.24) taken from Bauer (1983) and Lascarides and Copestake (1998). The discourse in (7.22a) favors the interpretation "man made of garbage" for *garbage man* (by analogy over snowman) over the more likely interpretation "man who collects garbage". Although *fast programmer* is typically a programmer who programs fast, when the adjective-noun combination is embedded in a context like (5.3a,b), repeated here as (7.23a,b), the less likely meaning "a programmer who runs fast" is triggered. Finally, although it is more likely to enjoy reading a book rather than eating it, the context in (7.24a) triggers the latter interpretation.

(7.22) a. In the back of the street where I grew up, everybody was poor. We were so poor that we never went on holiday. Our only toys were the garbage cans.

b. We never built sandcastles, only garbage men.

(7.23) a. All the office personnel took part in the company sports day last week.

b. One of the programmers was a good athlete, but the other was struggling to finish the courses.

c. The fast programmer came first in the 100m.

(7.24) a. My goat eats anything.

b. He really enjoyed your book.

The argument preferences which our model generates can be thought of as default semantic knowledge, to be used in the absence of any explicit contextual or lexical semantic information to the contrary. A model that takes pragmatic information into account would not only have to deal with data sparseness but furthermore detect cases where conflicts arise between discourse information and the likelihood of a given interpretation.

Not all misclassifications are the result of default interpretations being overridden by context. There are cases where the algorithm simply comes up with a counterintuitive analysis: in sentence (7.25a) the compound *animal construction* was given an object interpretation, and in (7.25b) *work provider* was given a subject interpretation.

(7.25) a. Coral polyps are each only a few millimetres across but working together in colonies, they have produced the greatest [animal$_{\mathrm{N}}$ constructions$_{\mathrm{N}}$] the world had seen before man began his labours.

b. You will be obliged to progress a case until you have formed a view on liability whether or not it has been vetted by the [work$_{\mathrm{N}}$ provider$_{\mathrm{N}}$].

Our experiments focused on nominalizations derived from verbs specifically subcategorizing for direct objects. Although nominalizations whose verbs take prepositional frames (e.g., *oil painting*, *soccer competition*) represent a small fraction of the nominalizations found in the corpus (9.2%), a more general approach would have to take them into account. This task is considerably harder since in order to estimate the frequency $f(v_{n_2}, rel, n_1)$, one needs to determine with some degree of accuracy the attachment site of the prepositional phrase first. Taking into account PPs and their attachment sites can be also useful for the interpretation of compounds other than nominalizations. Consider the compound noun *pet spray* from (7.1). Assuming that *pet spray* can be either "spray for pets", "spray in pets", "spray about pets", or "spray from pets", we can derive the most likely interpretation by looking at which types of PPs (e.g., *for pets*, *about pets*) are most likely to attach to *spray*. Note that in cases where the expressions *spray for pets* or *spray in pets* are not attested in the corpus their respective co-occurrence frequencies can be recreated using the techniques presented in Section 7.2.2. The approach advocated here can be straightforwardly extended to nominalizations with adjectival modifiers (e.g., *parental refusal*, see the examples in (7.2)). In most cases the adjective in question is derived from a noun and any inference process on the argument relations between the head noun and the adjectival modifier could take advantage of this information.

7.6. Related Work

In this section we review previous work on the interpretation of compound nouns. Despite their differences most approaches require large amounts of hand-crafted knowledge, place emphasis on the recovery of relations other than nominalizations (see the examples in (7.1)), and contain no qualitative evaluation (the exceptions are Leonard 1984, Vanderwende 1994, and Lauer 1995). Most symbolic approaches are limited to a specific domain due to the large effort involved in hand-coding semantic information and are distinguished in two main types: concept-based and rule-based.

Under the concept-based approach each noun is associated with a concept and various slots. Compound interpretation reduces to slot filling, i.e., evaluating how appropriate concepts

are as fillers of particular slots. A scoring system evaluates each possible interpretation and selects the highest scoring analysis. Examples of the approach are Finin (1980) and McDonald (1982). As no qualitative evaluation is reported it is difficult to assess how the method performs, although it is clear that considerable effort needs to be invested in the encoding of the appropriate semantic knowledge.

Under the rule-based approach interpretation is performed by sequential rule application. A fixed set of rules are applied in a fixed order, and the first rule for which the conditions are met results in the most plausible interpretation. The approach was introduced by Leonard (1984), was based on a hand-crafted lexicon, and achieved an accuracy of 76.0% (although on the training set). Vanderwende (1994) further developed a rule-based algorithm which no more relies on a hand-crafted lexicon, but extracts the required semantic information from an on-line dictionary instead. Vanderwende (1994) achieves an accuracy of 52.0%.

A variant of the concept-based approach uses unification to constrain the semantic relations between nouns represented as feature structures. Jones (1995) uses a typed graph-based unification formalism and default inheritance to specify features for nouns whose combination results in different interpretations. Again no evaluation is reported, although Jones (1995) points out that ambiguity can be a problem, as all possible interpretations are produced for a given compound. Wu (1993) provides a statistical framework for the unification-based approach and develops an algorithm for approximating the probabilities of different possible interpretations using the maximum entropy principle. No evaluation of the algorithm's performance is given. However, the approach still remains knowledge intensive as it requires manual construction of the feature structures.

Lauer (1995) provides a probabilistic model of compound noun paraphrasing (e.g., *state laws* are "the laws of the state", *war story* is "a story about war", etc.) which assigns probabilities to different paraphrases using a corpus in conjunction with Roget's publicly available thesaurus. Lauer (1995) does not address the interpretation of nominalizations or compounds with hyponymic relations (see example (7.1e)) and takes into account only prepositional paraphrases of compounds (e.g., *of*, *for*, *in*, *at*, etc.). Lauer's (1995) model makes predictions about the meaning of compound nouns on the basis of observations about prepositional phrases. The model combines the probability of the modifier given a certain preposition with the probability of the head given the same preposition, and assumes that these two probabilities are independent.

Consider for instance the compound *war story*. In order to derive the intended interpretation (i.e., "story about war") the model takes into account the frequency of *story about* and *about war*. The modifier and head noun are substituted by the concepts with which they are represented in Roget's thesaurus and the frequency of a concept and a preposition is calculated accordingly (see Section 7.2.2.2). Lauer's (1995) model achieves an accuracy of 47.0%. The result is difficult to interpret given that no experiments with humans are performed and

therefore the optimal performance on the task is unknown. Lauer (1995) acknowledges that data sparseness can be a problem for the estimation of the model parameters and also that the independence assumption between the head and its modifier is unrealistic and leads to errors in some cases.

Although it is generally acknowledged that context, both intra- and inter-sentential, may influence the interpretation task, contextual factors are typically ignored, with the exception of Hobbs et al. (1993) who propose that the interpretation of a compound can be achieved via abductive inference. In order to interpret a compound one must prove the logical form of its constituent parts from what is mutually known. However, the amount of world knowledge required to work out what is mutually known renders such an approach infeasible in practice. Furthermore, Hobbs et al.'s (1993) approach does not capture linguistic constraints on compound noun formation, and as a result cannot predict that a noun-noun sequence like *cancer lung* is odd.

Unlike previous work, we did not attempt to recover the semantic relations holding between a head and its modifier (see (7.1)). Instead we focused on the less ambitious task of interpreting nominalizations, i.e., compounds whose heads are derived from a verb and their modifiers are interpreted as its arguments. Similarly to Lauer (1995), we have proposed a simple probabilistic model which uses information about the distributional properties of words and domain independent symbolic knowledge (i.e., WordNet, Roget's thesaurus). Unlike Lauer (1995), we have addressed the sparse data problem by directly comparing and contrasting a variety of smoothing approaches proposed in the literature and have shown that these methods yield satisfactory results for the demanding task of semantic disambiguation. Furthermore, we have shown that the combination of different sources of taxonomic and non-taxonomic information (using decision tree learning) is effective for tasks facing data sparseness. Although the use of machine learning has been widespread in studies concerning discourse segmentation, the disambiguation of discourse clues, and the acquisition of lexical semantic classes (Hatzivassiloglou and McKeown 1995a, 1997; Hatzivassiloglou and Wiebe 2000; Merlo and Stevenson 1999; Passonneau and Litman 1997; Siegel 1999; Siegel and McKeown 1994), its application to the interpretation of compound nouns is novel.

Our approach can be easily adapted to account for Lauer's (1995) paraphrasing task. Instead of assuming that the probability of the compound modifier given a preposition is independent from the probability of the compound head given the same preposition, a more straightforward model would take into account the joint probability of the head, the preposition, and the modifier. In cases where a certain head, preposition, and modifier combination is not attested in the corpus (e.g., *story about war*), the methodology put forward in Experiments 16 and 17 could be used to recreate its frequency (see also the discussion in Section 7.5).

Unlike previous approaches, we provide an upper-bound for the task. Recall from Section 7.2.4 that an experiment with humans was performed so as to evaluate whether the task can

be performed reliably. In doing so we took context into account and as a result we established a higher upper bound for the task than would have been the case if context was not taken into account. Furthermore, it is not clear whether subjects could arrive at consistent interpretations for nominalizations out of context. Downing's (1977) experiments show that, when asked to interpret compounds out of context, subjects tend to come up with a variety of interpretations, which are not always compatible. For example, for the compound *bullet hole* the interpretations "a hole made by a bullet", "a hole shaped like a bullet", "a fast-moving hole", "a hole in which to hide bullets", and "a hole into which to throw (bullet) casings" were provided.

7.7. Summary

In this chapter we addressed the interpretation of nominalizations. We cast the interpretation task as a disambiguation problem and proposed a statistical model for inferring the argument relations holding between a deverbal head and its modifier.

We showed how the argument relations (which are not readily available in the corpus) can be retrieved by using partial parsing and smoothing techniques that exploit distributional and taxonomic information. We directly compared and contrasted a variety of smoothing approaches proposed in the literature and demonstrated that these methods yield satisfactory results for the demanding task of semantic disambiguation. Our approach is applicable to domain independent unrestricted text and does not require the hand coding of semantic information.

In Chapter 6 we demonstrated how the surface cueing approach performs under data sparseness with regard to the acquisition of compound nouns, an extremely productive grammatical process. In this chapter we showed how data sparseness affects the probabilistic interpretation of compound nouns. In both cases we compensated for the lack of sufficient distributional information using methods that either *directly* "recreate" the frequencies of word combinations or features whose distribution in the corpus *indirectly* provides information about the likelihood of occurrence of rare word combinations. In both cases we used machine learning to combine different sources of taxonomic and non-taxonomic information.

In the case of acquisition, we showed that non-taxonomic features such as the frequency of the compound head, the likelihood of a word as a modifier, or the context surrounding a candidate compound, perform almost as well as features that are estimated on the basis of existing taxonomies such as WordNet (e.g., the frequency of the concepts representing the candidate compound). As far as the interpretation is concerned, we showed that good performance can be achieved by taking advantage of smoothing methods that rely simply on verb-argument tuples extracted from a large corpus (e.g., back-off and Jensen). For both tasks, however, the best performance was achieved when both types of information (i.e., taxonomic and non-taxonomic) were taken into account.

Chapter 8

Conclusions

This chapter summarizes the main findings of this thesis and outlines some issues for further research.

8.1. Main Findings

This thesis investigated the acquisition and probabilistic modeling of lexical knowledge. We focused on systematic polysemy, i.e., the regular and predictable meaning alternations to which certain classes of words are subject, and presented a series of empirical studies that explored the systematic correspondences between meaning and syntax in order to discover polysemous lexical units and disambiguate their meaning. The results of this investigation provided a series of experimental, methodological and theoretical contributions with regard to the formalization, acquisition, and modeling of systematic polysemy. The following is a summary of the central findings.

Linguistic theory. We conducted a series of experiments that used insights from linguistic theory in order to guide and structure the acquisition of systematically polysemous units from domain independent wide-coverage text. These experiments focused on three major syntactic relations (verbs and their complements, adjective-noun combinations, and noun-noun modification) and exploited the correspondence between meaning and syntax in the acquisition process.

Our experimental results demonstrated that the framework of a linguistic theory can be effectively combined with induction methods from data-intensive linguistics and machine learning. Unlike purely corpus-based analyses, the framework of a semantic theory provides generalizations about the behavior of lexical units. These generalizations and their predictions can be empirically tested using corpus-based measures. Employing a semantic theory presupposes a richer representation of lexical meaning and allows for the acquisition of refined semantic features that do not emerge from purely statistical collocational analysis.

Our findings showed that corpus data can be used to discover novel facts about the behavior of lexical units that are not readily available from linguistic introspection. The information acquired from the corpus (e.g., corpus-derived meanings, subcategorization frame frequencies, productivity and typicality estimates) can be used to quantify and enrich linguistic theory.

Probabilistic modeling. We proposed a probabilistic model which exploits corpus-based distributions of lexical information to select the most dominant (i.e., salient) meaning from a set of meanings for systematically polysemous lexical units without taking discourse context into account. We modeled default meaning—the most likely meaning for a given word combination across all of its discourse contexts—probabilistically in Bayesian framework which combines observed linguistic dependencies (in the form of conditional probabilities) with linguistic generalizations (in the form of prior probabilities derived directly from the corpus or from classifications such as Levin 1993).

We explored the generality of the model by applying it to verbs and their complements, noun-noun compounds, and adjective-noun combinations. Besides providing a probabilistic formalization of systematic polysemy, we presented a series of experiments that show how the various model parameters can be estimated using a combination of corpus counts and linguistic knowledge, either in the form of linguistic generalizations (e.g., Levin 1993, Levi 1978) or taxonomies (e.g., WordNet).

Our modeling studies focused on semantic ambiguity arising from word combinations as opposed to individual words. We presented several experiments which explored the appropriate meaning inventory for the task and showed that the model performs adequately under different assumptions with respect to the type and number of available meanings. In Chapter 4 we used Levin's (1993) semantic classes to describe the meanings of verbs with regard to their complements. In Chapter 5 the meanings of adjective-noun combinations were approximated by verbs which are modified by the adjective and subcategorize for the noun present in the adjective-noun pair. In Chapter 7 the meanings were simply argument relations.

Our findings showed that the model is general enough to account for different types of lexical units under varying assumptions about data requirements (sufficient versus sparse data) and meaning representations (corpus internal versus corpus external).

Surface cueing. We demonstrated that the surface cueing approach performs reasonably well at obtaining information pertaining to lexical units. The experiments in this thesis revealed that a combination of shallow parsing and linguistically motivated heuristics can be effectively applied both for the acquisition and interpretation of lexical semantic information.

Our modeling and acquisition studies explored how the proposed approach fared under data sparseness. In particular, we investigated how the surface cueing approach can be adapted to acquire and disambiguate lexical units for which very little evidence is found in the corpus.

Our experiments in Chapters 6 and 7 showed that the lack of sufficient distributional information can be compensated by using either methods that *directly* "recreate" the frequencies of word combinations (disambiguation task), or features whose distribution in the corpus *indirectly* provides cues about the likelihood of the occurrence of rare word combinations (acquisition task). In both cases we showed that the combination of a variety of information sources performs reliably better than each source individually, employing decision tree learning as a uniform methodology.

The findings of this thesis are potentially of interest both to linguistic theory and practical NLP applications. We provide a relatively straightforward methodology for acquiring, testing, and quantifying linguistic generalizations. The acquired corpus distributions are explored in a probabilistic framework which constrains lexical ambiguity by providing a ranking of alternative interpretations. The latter can be of benefit to applications such as machine translation, information retrieval/extraction, and natural language generation.

In summary, this thesis contributed a probabilistic model of systematic polysemy. This model is grounded in linguistic theory as it exploits linguistic insights both in the formalization of polysemy and the acquisition of the model parameters. We demonstrated how this model can be used for the resolution of semantic ambiguity using a heuristic approach which relies on the shallow processing of corpora and the availability of domain independent lexical resources such as WordNet.

8.2. Issues for Further Research

In this section, we provide a brief discussion of a number of issues for further research that follow from the findings reported in this thesis.

8.2.1. Further Modeling and Acquisition Studies

A number of additional modeling and acquisition studies can be carried out to test the generality of the proposed approach. As modeling and acquisition are methodologically interrelated we discuss extensions to both tasks in this section.

The experiments in Chapters 3 and 4 focused on a limited number of alternations and their related frames. Experiments with a full-scale subcategorization dictionary could provide important information not only about a particular alternation but also about its properties in relation to other alternations. For example, experimentation with frames pertaining to a variety of alternations could provide answers to the following questions: (a) what types of alternations are most productive (e.g., whether transitivity alternations are more productive than reflexive diathesis alternations, see Section 3.1 for details on the different types of alternations), (b) how widespread are alternations in corpus data; in other words, what is the relative proportion of typical versus atypical verbs across all alternations, and (c) what linguistic information is needed

to acquire frames relating to alternations; note that in our studies we made use of shallow syntactic, grammatical and semantic information in order to acquire corpus tokens faithful to the meaning and structure of frame alternants. Morphological or argument structure, i.e., thematic role, information may prove crucial for the acquisition of certain types of alternations.

Experimentation with a full-scale subcategorization dictionary would be also useful for addressing questions concerned with the modeling of verb class ambiguity (see Chapter 4). In particular, we would able to answer the following: (a) What are the frames for which the semantic class is predicted most accurately? (b) What are the verbs for which the semantic class is predicted most accurately? (c) How does ambiguity influence the performance of the model, i.e., does the model perform better with verbs which exhibit a relatively low level of ambiguity? Furthermore, the acquisition of a subcategorization dictionary would enable a more realistic estimation of the model parameters. Recall from Chapter 4 that one of the parameters of our model is the frame probability $P(f)$. In our experiments we relied on Levin (1993) to provide an estimate of this value. Alternatively, we could have acquired Levin compatible subcategorization frames from the BNC. Another issue related to the model presented in Chapter 4 is the type of information sources used to predict the most dominant verb semantic class. Our model relied solely on Levin's (1993) classification and subcategorization frames acquired from the BNC. It is an interesting empirical question as to whether the performance of the model could be improved by incorporating additional sources of information such as local context, argument structure, and selectional restrictions.

In Chapter 5 we focused on polysemous adjectives whose meaning varies depending on the noun they modify. The phenomenon is also observed with verbs where different meaning seems to arise depending on their complements (e.g., *Ann enjoyed the book* versus *Ann enjoyed the ice-cream*, see also the examples in (2.11)). A further extension to the model presented in Chapter 5 would be to account for the meaning of these verbs. This would also enable us to observe whether there are any differences in model predictions with respect to adjectives and verbs. For example, it may be the case that (all else being equal) a higher correlation is found between the model's rankings and human judgments for verbs than for adjectives, since verbs tend to impose stricter semantic restrictions on the nouns with which they combine than adjectives. Or it may be the case that the number of meanings found for verbs is smaller than those found for adjectives.

Other extensions concern the acquisition and interpretation of compound nouns. Our experiments in Chapter 6 dealt solely with noun-noun compounds. An interesting question is whether the methodology introduced in this chapter can be adapted to account for less productive types of compounds such as adjective-noun compounds (e.g., *heavy metal*) and compounds with possessives as modifiers of the head noun (e.g., *cashier's check*, see Section 6.1 for an overview of the different types of compounds). Another interesting issue concerns the acquisition of compounds consisting of more than two words. Part of this process involves the

analysis of the candidate compound in its subcomponents (e.g., *physical therapy program* can be bracketed as [[*physical therapy*] *program*]). Lauer's (1995) work on the analysis of compound nouns concentrated only on three word compounds consisting solely of sequences of nouns. Lauer proposed a probabilistic model which infers the bracketing on the basis of Roget's thesaurus and corpus frequencies. The combinatory approach put forward in Chapter 6, where different information sources are exploited to make predictions about the status of candidate compounds, could be extended to the analysis and acquisition of candidates of length longer than two containing not only nouns but also adjectives (e.g., *physical therapy program*).

As far as the modeling of the meaning of compound nouns is concerned, our experiments in Chapter 7 addressed only noun-noun nominalizations. Obvious extensions include nominalizations with adjectival modifiers (e.g., *parental refusal*) and nominalizations of length larger than two (e.g., *council tax administration*). The results of Chapter 7 indicated that smoothing methods can satisfactorily recreate the missing frequencies of syntactic relations. A similar methodology could be employed for the interpretation of compounds other than nominalizations (see Section 7.1 for details on the semantic relations between the compound head and its modifier).

Finally, note that systematic polysemy applies to a variety of nouns other than compounds (see Apresjan 1973 for an overview of the systematic meaning alternations for nouns). Recall from Chapter 1 (Section 1.2.1) that count nouns can be extended to mass nouns, containers can be extended to their contents, etc. A further application for the proposed model would be to predict meaning preferences for systematic polysemous nouns. This would involve further research on the surface cueing approach since cues indicating the different meanings of the noun would have to be gleaned from the corpus. For example, the noun *rabbit* is more likely to take on the "animal" meaning when it is attested in plural form or when it is modified by a determiner (e.g., *Rabbits are great travelers, I want to see the rabbit*). The "meat" meaning is more likely when *rabbit* is a bare singular noun in an object position (e.g., *He buys rabbit from the local butcher*).

8.2.2. Methodological Issues

In the experiments reported throughout this thesis we put forward an approach which favors the shallow syntactic analysis of the corpus. Although the approach delivers satisfactory results, it is a matter of future research to explore whether performance differences are observed with a parser that attempts the complete analysis of text instead of focusing on basic syntactic relations. Recent advances in large-scale, grammar-based parsing have led to the development of efficient parsing algorithms that utilize wide-coverage, general-purpose grammars (see Carroll and Oepen 2000 for an overview). Statistical models for parsing natural language have also shown considerable successes in broad-coverage domains (e.g., Carroll and Rooth 1998, Collins 1998). With respect to the experiments reported in this thesis, the output of a large-scale

parsing system could be used either for the acquisition of subcategorization frames or for the extraction of argument relations (e.g., verb-subject, verb-object).

The proposed model took only minimal context into account in order to infer the dominant meaning for a given word combination. Further experiments could investigate the influence of a larger context in the modeling of dominant meanings. Another issue concerns the different ways contextual information is encoded. For example, context can be simply represented as the words surrounding the disambiguation target, or their parts of speech, or both. Context can be also represented as syntactic relations or semantic categories. Future work would have to assess the impact of these choices in the modeling process.

The majority of the experiments in this thesis were conducted on the BNC, a wide-coverage synchronic corpus of British English. We did not investigate how corpus size and corpus register influence the obtained results (see Roland and Jurafsky 2000 for discussion on the effects of corpus size and register on corpus-derived frequencies). We would expect the ranking of alternative interpretations to vary across domains. Consider a hypothetical corpus of cooking recipes. In this context we would expect the "meat" meaning to be more likely for *rabbit* than the "animal" meaning. Also, an adjective-noun combination such as *easy flour* would be more likely to receive the interpretation "flour that is easy to bake" or "flour that rises easily" instead of "flour that is easy to buy" or "flour that is easy to produce". Another related issue for further investigation is how corpus size influences the ranking of the available meanings or even the size of the meaning inventory in cases where this is derived from the corpus.

8.2.3. The Lexicon

In this thesis we have argued that the acquired probabilistic information can extend the empirical base of linguistic theory and reduce the proliferation of lexical semantic ambiguity that is inherent in linguistic generalizations. We have not, however, addressed the issue of combining probabilistic information with lexicon formalisms in novel and useful ways.

Further research has to look into ways of incorporating frequency information into the lexicon. There are a number of questions concerning the structure and design of the lexicon when frequency data is taken into account. Existing theories of the lexicon, such as the generative lexicon (Pustejovsky 1995) and unification/constraint-based accounts of the lexicon (e.g., Copestake 1992, Pollard and Sag 1994) cannot readily integrate frequency data. We demonstrated in this thesis how frequency data corresponding to lexical generalizations can be acquired. It remains to be shown how this frequency data interacts with an inheritance-based view of the lexicon. It is not clear where such frequency data must be stated, or how it might be inherited. In this thesis we have derived frequency information both for individual lexical items and the semantic classes they correspond to. It is conceivable that frequency information is associated directly with individual lexical entries or with rules that apply to classes of lexical

entries (see Briscoe and Copestake 1999 for some discussion). A related question concerns the interplay between those two types of information. In particular, more work is needed to determine when a lexical item should be assumed to correspond to the empirical behavior of its semantic class and when it should be treated as a special case.

8.2.4. Word Sense Disambiguation

Throughout this thesis we have argued in favor of a task related to and yet distinct from word sense disambiguation. We presented a model that delivers the most likely meaning out of a set of meanings across all discourse contexts, rather than choosing the appropriate meaning in a particular discourse context. We have not, however, systematically explored how our proposal can be combined with recent advances in word sense disambiguation.

Two important questions arise for future research. The first question involves the meaning inventory employed in this thesis and the phenomena under consideration. As mentioned in Chapter 1 systematic polysemy has not been the main focus of work in word sense disambiguation. Our experimental results suggest novel directions for word sense disambiguation research where emphasis is placed not only on single words but on systematically polysemous lexical units. The second question concerns the potential usefulness of our approach for word sense disambiguation methods. In particular, our dominant meanings can be used either to produce the training data for supervised methods (using bootstrapping) or to form the basis of novel unsupervised approaches to word sense disambiguation which combine context specific cues with prior context invariant information about the likelihood of the meanings of a given word or word combination.

8.2.5. Semantic Defaults and Intuitions

Although the work reported in this thesis is not psycholinguistic, it offers interesting possibilities for psycholinguistic research. Our evaluation studies in Chapter 5 demonstrated that the derived meanings and their ranking are reliably correlated with human judgments. A range of further studies can be conducted to assess the degree to which the derived meaning preferences correspond to human intuitions. Under the assumption that large corpora provide a useful sample of the language to which people are exposed, corpus-based studies such as the ones presented in this thesis can potentially shed light on the acquisition of semantic knowledge by humans.

Future work in this direction involves determining the degree to which the corpus-based typicality and productivity values (see Chapter 3) correspond to preferences humans have with respect to alternating verbs. Our model's interpretation preferences for nominalizations can be also compared against preferences obtained from humans. An interesting question is whether people can agree on their interpretations without taking discourse context into ac-

count. A variety of psycholinguistic studies have supported a cognitively relevant role for the statistical knowledge inherent in the linguistic environment (Howes and Solomon 1951; Monsell 1991; Redington, Chater, and Finch 1998; Saffran, Aslin, and Newport 1996; Schooler and Anderson 1997; Spence and Owens 1990; Whaley 1978). Recent work in computational psycholinguistics has exploited the distributional information present in large language corpora in order to model behavioral data (Jurafsky 1996; Lapata et al. 2001; McDonald 2000; Narayanan and Jurafsky 1998; Resnik and Diab 2000; Resnik 1993). Future work in this area could assess the extent to which our model of dominant meanings quantifies native speaker's intuitions about the semantic properties of ambiguous words.

Appendix A

Annotation Guidelines

A.1. Verb Class Ambiguity

You will be given a series of sentences extracted automatically from the British National Corpus. In this experiment we are interested in verbs and their meanings and your task will be to associate these verbs with a list of semantic classes. In what follows you are given a description of these classes and examples illustrating which classes are appropriate for different verbs. Please read the instructions carefully before proceeding with the annotation.

convey (NP1 V NP2 *to* NP3) You will annotate three classes for the verb *convey*: SAY (SAY), SEND (SEND), and OTHER (O). Examples are given as follows:

(A.1) a. Jane conveyed the message to her sister. SAY
b. Mary conveyed her sympathy to my parents. SEND
c. Mary conveyed the news to the extent. O

- Use the class SAY when we convey something to someone by speaking (e.g., messages, news) (see (A.1a)).
- Use the class SEND when the means of conveying something to someone remains underspecified (e.g., appreciation, impression) (see (A.1b)).
- When there is ambiguity between the SEND reading and SAY reading default to the latter.
- Assign the class OTHER to tokens which have either the wrong frame or for which the two classes are not applicable (see (A.1c)). Also, assign the class OTHER if the verb is followed by a particle (i.e., if it is a phrasal verb).

express (NP1 V NP2 *to* NP3) You will annotate three classes for the verb *express*: SEND (SEND), REFLEXIVE VERBS OF APPEARANCE (RAP), and OTHER (O). Examples are given as follows:

(A.2) a. Peter expressed his gratitude to Frank. SEND
b. Mary expressed herself to the judge. RAP
c. Mary expressed her devotion to her friends. O

- Use the class SEND when we express something to someone (see (A.2a)).
- Use the class REFLEXIVE VERBS OF APPEARANCE when the NP object of the verb is a reflexive pronoun and agrees with the subject in number and gender (see (A.2b)).
- Assign the class OTHER to tokens which have either the wrong frame or for which the two classes are not applicable (see (A.2c)). Also, assign the class OTHER if the verb is followed by a particle (i.e., if it is a phrasal verb).

extend (NP1 V NP2 *to* NP3) You will annotate three classes for the verb *extend*: CONTRIBUTE (CON), FUTURE HAVING (FH), and OTHER (O). Examples are given as follows:

(A.3) a. Jane extended her gratitude to her boss. FH
b. They extended the party to the garden. CON
c. Peter extended his property to few. O

- Use the class FUTURE HAVING when the object of the preposition *to* is animate or denotes a group (e.g. association, village) (see (A.3a)).
- Use the class CONTRIBUTE when the object of the preposition is not animate (see (A.3b)).
- Assign the class OTHER to tokens which have either the wrong frame or for which the two classes are not applicable (see (A.3c)). Also, assign the class OTHER if the verb is followed by a particle (i.e., if it is a phrasal verb).

fly (NP1 V NP2 *to* NP3) You will annotate three classes for the verb *fly*: DRIVE (DR), RUN (RUN), and OTHER (O). Examples are given as follows:

(A.4) a. John flew Frank to Kosovo. DR
b. Frank flew the plane to Kosovo. RUN
c. Frank flew south to Stuttgart O

- Use the class DRIVE when the dative object is animate. Basically when we fly someone somewhere (see (A.4a)).

- Use the RUN class when we fly something somewhere (see (A.4b)).
- Assign the class OTHER to tokens which have either the wrong frame or for which the two classes are not applicable (see (A.4c)). Also, assign the class OTHER if the verb is followed by a particle (i.e., if it is a phrasal verb).

issue (NP1 V NP2 *to* NP3) You will annotate three classes for the verb *issue*: FUTURE HAVING (FH), FULFILLING (FUL), and OTHER (O). Examples are given as follows:

(A.5) a. I issued a ticket to John. FH
b. The leader issued a warning to his followers. FUL
c. Mary issued a ticket to Norway. O

- Use the class FUTURE HAVING for actual things we issue to people (e.g., contracts, certificates, tickets, invoices, etc.) (see (A.5a)).
- Use the class FULFILLING for more abstract concepts (see (A.5b)).
- Assign the class OTHER to tokens which have either the wrong frame or for which the two classes are not applicable (see (A.5c)). Also, assign the class OTHER if the verb is followed by a particle (i.e., if it is a phrasal verb).

leave (NP1 V NP2 *to* NP3) You will annotate three classes for the verb *leave*: FULFILLING (FUL), FUTURE HAVING (FH), and OTHER (O). Examples are given as follows:

(A.6) a. My grandmother left a lot of money to the poor. FH
b. We will leave the matter to John. FUL
c. We left it to maximum. O

- Use the class FUTURE HAVING when the verb refers to a change of possession (see (A.6a)).
- Use the class FULFILLING when we are leaving something to someone/something (see (A.6b)).
- When a sentence is ambiguous between the FULFILLING and FUTURE HAVING reading default to the FULFILLING class.
- Assign the class OTHER to tokens which have either the wrong frame or for which the two classes are not applicable (see (A.6c)). Also, assign the class OTHER if the verb is followed by a particle (i.e., if it is a phrasal verb).

offer (NP1 V NP2 *to* NP3) You will annotate three classes for the verb *offer*: FUTURE HAVING (FH), REFLEXIVE VERBS OF APPEARANCE (RAP), and OTHER (O). Examples are given as follows:

(A.7) a. John offered a big ice-cream to Jane. FH
b. The solution offered itself to Peter. RAP
c. Peter offered an introduction to syntax. O

- Use the class FUTURE HAVING when we offer something to someone (see (A.7a)).
- Use the class REFLEXIVE VERBS OF APPEARANCE when the NP object of the verb is a reflexive pronoun and agrees with the subject in number and gender (see (A.7b)).
- Assign the class OTHER to tokens which have either the wrong frame or for which the two classes are not applicable (see (A.7c)). Also, assign the class OTHER if the verb is followed by a particle (i.e., if it is a phrasal verb).

pass (NP1 V NP2 *to* NP3) You will annotate four classes for the verb *pass*: GIVE (GIVE), SEND (SEND), THROWING (THR), and OTHER (O). Examples are given as follows:

(A.8) a. Frank passed the book to Jane. GIVE
b. I passed the complaint to my superiors. SEND
c. Jane passed the ball to Peter. THR
d. I pass my exams every year. O

- Use the class GIVE when we are passing something to someone. In particular, when we are passing something which is not an abstract entity and the person who receives it will be its possessor (see (A.8a)).
- Use the class SEND when we are passing something which denotes either a state, or an abstract entity and the person who receives it is not necessarily its possessor (see (A.8b)).
- Use the class THROWING when pragmatic knowledge indicates that the object we are passing is being thrown (see (A.8c))
- In case of ambiguity between the SEND and GIVE reading default to the GIVE class.
- Assign the class OTHER to tokens which have either the wrong frame or for which the two classes are not applicable (see (A.8d)). Assign the class OTHER if the verb is followed by a particle (i.e., if it is a phrasal verb).

pose (NP1 V NP2 *to* NP3) You will annotate three classes for the verb *pose*: MESSAGE TRANSFER (MT), REFLEXIVE VERBS OF APPEARANCE (RAP), and OTHER (O). Examples are given as follows:

(A.9) a. The competition poses a great challenge to all PhD students. MT
b. A great danger posed itself to Peter yesterday. RAP
c. Frank poses solutions to every problem. O

- Use the class MESSAGE TRANSFER when we pose something to someone (see (A.9a)).
- Use the class REFLEXIVE VERBS OF APPEARANCE when the NP object of the verb is a reflexive pronoun and agrees with the subject in number and gender (see (A.9b)).
- Assign the class OTHER to tokens which have either the wrong frame or for which the two classes are not applicable (see (A.9c)). Also, assign the class OTHER if the verb is followed by a particle (i.e., if it is a phrasal verb).

present (NP1 V NP2 *to* NP3) You will annotate three classes for the verb *present*: FULFILLING (FUL), REFLEXIVE VERBS OF APPEARANCE (RAP), and OTHER (O). Examples are given as follows:

(A.10) a. John presented his book to the audience. FUL
b. Dolores presented herself to her students. RAP
c. Frank presented his masterpiece. O

- Use the class FULFILLING when we present something to someone (see (A.10a)).
- Use the class REFLEXIVE VERBS OF APPEARANCE when we present ourselves to someone. In other words, when the dative object is a reflexive pronoun (see (A.10b)).
- Assign the class OTHER to tokens which have either the wrong frame or for which the two classes are not applicable (see (A.10c)). Also, assign the class OTHER if the verb is followed by a particle (i.e., if it is a phrasal verb).

return (NP1 V NP2 *to* NP3) You will annotate three classes for the verb *return*: CONTRIBUTE (CON), SEND (SEND) and OTHER (O). Examples are given as follows:

(A.11) a. Nato returned the refugees to Kosovo. CON
b. Peter returned the books to my sister. SEND
c. I returned the books to my satisfaction. O

- Use the class CONTRIBUTE when the object of the preposition is not animate (see (A.11a)).

- Use the class SEND when the object of the preposition *to* is animate or denotes a group (e.g., association, village, etc.) (see (A.11b)).
- Assign the class OTHER to tokens which have either the wrong frame or for which the two classes are not applicable (see (A.11c)). Also, assign the class OTHER if the verb is followed by a particle (i.e., if it is a phrasal verb).

serve (NP1 V NP2 *to* NP3) You will annotate three classes for the verb *serve*: GIVE (GIVE), FULFILLING (FUL), and OTHER (O). Examples are given as follows:

(A.12) a. I served my mother her dinner. GIVE
b. John served his time to the governor. FUL
c. Mary serves Monday to Friday. O

- Use the class GIVE for actual things we serve to people (see (A.12a)).
- Use the class FULFILLING for more abstract concepts (see (A.12b)).
- Assign the class OTHER to tokens which have either the wrong frame or for which the two classes are not applicable (see (A.12c)). Also, assign the class OTHER if the verb is followed by a particle (i.e., if it is a phrasal verb).

show (NP1 V NP2 *to* NP3) You will annotate three classes for the verb *show*: MESSAGE TRANSFER (MT), REFLEXIVE VERBS OF APPEARANCE (RAP), and OTHER (O). Examples are given as follows:

(A.13) a. I showed my dress to Jane. MT
b. The film showed itself to Peter yesterday. RAP
c. Frank finally showed his addictions to drugs. O

- Use the class MESSAGE TRANSFER when we show something to someone. In most cases the object of the preposition *to* is animate (see (A.13a)).
- Use the class REFLEXIVE VERBS OF APPEARANCE when the NP object of the verb is a reflexive pronoun and agrees with the subject in number and gender (see (A.13b)).
- Assign the class OTHER to tokens which have either the wrong frame or for which the two classes are not applicable. In some cases you will find out that the PP can attach either to verb or to the noun. In this case, use your intuitions to decide whether the PP is OTHER or any of the other two classes (see (A.13c)). Also, assign the class OTHER if the verb is followed by a particle (i.e., if it is a phrasal verb).

suggest (NP1 V NP2 *to* NP3) You will annotate three classes for the verb *suggest*: REFLEXIVE VERBS OF APPEARANCE (RAP), SAY (SAY), and OTHER (O). Examples are given as follows:

(A.14) a. The solution suggested itself to Mary. RAP
b. Peter suggested the changes to his boss. SAY
c. John suggested an addiction to drugs. O

- Use the class REFLEXIVE VERBS OF APPEARANCE when something suggests itself to someone. In other words, when the dative object is a reflexive pronoun and agrees in number and gender with the subject (see (A.14a)).
- Use the class SAY when we suggest something to someone (see (A.14b)).
- Assign the class OTHER to tokens which have either the wrong frame or for which the two classes are not applicable (see (A.14c). Also, assign the class OTHER if the verb is followed by a particle (i.e., if it is a phrasal verb).

tell (NP1 V NP2 *to* NP3) You will annotate three classes for the verb *tell*: MESSAGE TRANSFER (MT), TELL (TL), and OTHER (O). Examples are given as follows:

(A.15) a. Pete told the secret to his brother. MT
b. Chris tells lies to my face. TL
c. Beth told her brother to book. O

- Use the class MESSAGE TRANSFER when the recipient is animate (see (A.15a)).
- Use the class TELL when the recipient is not animate (see (A.15b)).
- Assign the class OTHER when the recipient is not a prepositional phrase (see (A.15c)). Also, assign the class OTHER if the verb is followed by a particle (i.e., if it is a phrasal verb).

transfer (NP1 V NP2 *to* NP3) You will annotate three classes for the verb *transfer*: CONTRIBUTE (CON), SEND (SEND), and OTHER (O). Examples are given as follows:

(A.16) a. John transferred the problem to Mary. SEND
b. Frank transferred the subjects to the lab. CON
c. Mary transferred the asset to use. O

- Use the CONTRIBUTE when the object of the preposition is not animate (see (A.16b)).
- Use the SEND class when the object of the preposition *to* is animate or denotes a group (i.e., department, association, etc.) (see (A.16a)).

- Assign the class OTHER to tokens which have either the wrong frame or for which the two classes are not applicable (see (A.16c)). Also, assign the class OTHER if the verb is followed by a particle (i.e., if it is a phrasal verb).

write (NP1 V NP2 NP3) You will annotate three classes for the verb *write*: MESSAGE TRANSFER (MT), PERFORMANCE (PER), and OTHER (O). Examples are given as follows:

(A.17) a. Mary wrote Jane a note. MT
b. John wrote himself a book. PER
c. I wrote the book. O

- Use the class MESSAGE TRANSFER when we write things to someone, i.e., when there is a recipient who will receive the things we have written (see (A.17a)).
- Use the class PERFORMANCE when we write things for someone, i.e., in their favor. Also, assign the class PERFORMANCE when the dative object is a reflexive pronoun (see (A.17b)).
- When there is ambiguity between the PERFORMANCE reading and the MESSAGE TRANSFER reading default to the latter. In general, when in doubt, use pragmatic knowledge: we write a book for someone but not to someone.
- Assign the class OTHER to tokens which have a frame other than the double object (see (A.17c)). Also, assign the class OTHER if the verb is followed by a particle (i.e., if it is a phrasal verb).

call (NP1 V NP2 NP3) You will annotate three classes for the verb *take*: DUB (DUB), GET (GET), and OTHER (O). Examples are given as follows:

(A.18) a. Mary called him a fool. DUB
b. I will call you a taxi. GET
c. John called his mother every minute. O

- Use the class DUB when the subject is assigned a property or an attribute (see (A.18a)).
- Use the class GET when you want to get something or someone by calling them (see (A.18b)).
- Assign the class OTHER to tokens which have either the wrong frame or for which the two classes are not applicable (see (A.18c)). Also, assign the class OTHER if the verb is followed by a particle (i.e., if it is a phrasal verb).

cook (NP1 V NP2 NP3) You will annotate three classes for the verb *cook*: BUILD (BUILD), PREPARING (PREP), and OTHER (O). Examples are given as follows:

(A.19) a. Mary cooked Frank a wonderful meal. BUILD
b. Frank cooked me some eggs. PREP
c. I cook eggs every Sunday. O

- Use the class BUILD when the we are cooking a meal (e.g., breakfast, dinner) for someone (see (A.19a)). In this case the meal comes into existence.
- In all other cases, when we are cooking something for someone which already exists, use the class PREPARING (see (A.19b)).
- Assign the class OTHER to tokens which have either the wrong frame or for which the two classes are not applicable (see (A.19c)). Also, assign the class OTHER if the verb is followed by a particle (i.e., if it is a phrasal verb).

declare (NP1 V NP2 NP3) You will annotate three classes for the verb *declare*: DECLARE (DEC), REFLEXIVE VERBS OF APPEARANCE (RAP), and OTHER (O). Examples are given as follows:

(A.20) a. The president declared Frank press secretary. DEC
b. I declared myself a princess. RAP
c. Kate declared war. O

- Use the DECLARE class when we declare something or someone other than ourselves. In other words, when we attribute someone or something a property/characteristic (see (A.20a)).
- Use the class REFLEXIVE VERBS OF APPEARANCE when we attribute ourselves a property/characteristic. In other words, when the dative object is a reflexive pronoun (see (A.20b)).
- Assign the class OTHER to tokens which have either the wrong frame or for which the two classes are not applicable (see (A.20c)). Also, assign the class OTHER if the verb is followed by a particle (i.e., if it is a phrasal verb).

feed (NP1 V NP2 NP3) You will annotate three classes for the verb *feed*: FEEDING (FEED), GIVE (GIVE), and OTHER (O). Examples are given as follows:

(A.21) a. Peter fed Mary two bars of chocolate. FEED
b. Jane fed me many complaints. GIVE
c. Mary fed you the first time. O

- Use the class FEEDING when we feed someone food (see (A.21a)).
- Use the class GIVE when we feed someone with things which are not food (see (A.21b)).
- When a sentence is ambiguous between the FEEDING and GIVE reading default to the GIVE class.
- Assign the class OTHER to tokens which have either the wrong frame or for which the two classes are not applicable (see (A.21c)). Also, assign the class OTHER if the verb is followed by a particle (i.e., if it is a phrasal verb).

find (NP1 V NP2 NP3) You will annotate three classes for the verb *find*: DECLARE (DEC), GET (GET), and OTHER (O). Examples are given as follows:

(A.22) a. I find Harry stupid. DEC
b. Peter found me a taxi the other day. GET
c. I find money every month. O

- Use the class DECLARE when you assign someone a property or an attribute (see (A.22a)).
- Use the class GET when you want to get something or someone by finding them (see (A.22b)).
- Assign the class OTHER to tokens which have either the wrong frame or for which the two classes are not applicable (see (A.22c)). Also, assign the class OTHER if the verb is followed by a particle (i.e., if it is a phrasal verb).

leave (NP1 V NP2 NP3) You will annotate four classes for the verb *leave*: FUTURE HAVING (FH), FULFILLING (FUL), GET (GET), and OTHER (O). Examples are given as follows:

(A.23) a. I will leave myself some money. GET
b. Peter left me many debts. FH
c. John left Mary a widow. FUL
d. Johns leaves the office every week. O

- Use the class GET when the dative object is a reflexive pronoun and when we are leaving something for someone (see (A.23a)).
- Use the class FUTURE HAVING when the verb refers to a change of possession that will take place in the future (see (A.23b)). Use the class FUTURE HAVING when the verb can alternate in the prepositional frame (NP1 V NP2 *to* NP3).

- Use the class FULFILLING when the verb does not alternate in the prepositional frame (NP1 V NP2 *to* NP3) (see (A.23c)).
- Assign the class OTHER to tokens which have either the wrong frame or for which the two classes are not applicable (see (A.23d)). Also, assign the class OTHER if the verb is followed by a particle (i.e., if it is a phrasal verb).

make (NP1 V NP2 NP3) You will annotate three classes for the verb *make*: DUB (DUB), BUILD (BUILD), and OTHER (O). Examples are given as follows:

(A.24) a. I made John a captain. DUB
b. Frank made me a cup of tea. BUILD
c. I made a sandwich the same morning. O

- Use the class DUB when you assign someone or something a property or an attribute (see (A.24a)).
- Use the class BUILD when we are bringing something into existence (see (A.24b)).
- Also, use the class BUILD when we are doing something for someone or when the dative object is a reflexive pronoun.
- Assign the class OTHER to tokens which have either the wrong frame or for which the two classes are not applicable (see (A.24c)). Also, assign the class OTHER if the verb is followed by a particle (i.e., if it is a phrasal verb).

pass (NP1 V NP2 NP3) You will annotate four classes for the verb *pass*: GIVE (GIVE), SEND (SEND), THROWING (THR), and OTHER (O). Examples are given as follows:

(A.25) a. John passed you the salt. GIVE
b. Mary passed me your greetings. SEND
c. Jane suddenly passed me the ball. THR
d. I pass the train station every day. O

- Use the class GIVE when the non-dative object is usually a non-animate, non-abstract object (see (A.25a)).
- In some cases the class GIVE can be used when the non-dative object is animate (e.g., *Pass me Frank on the phone*).
- Use the class SEND when the non-dative object is an abstract concept (see (A.25b)).
- Use the class THROWING when pragmatic knowledge indicates that the object we are passing is being thrown (see (A.25c)).

- When there is ambiguity between the THROWING and GIVE reading default to the latter.
- Assign the class OTHER to tokens which have either the wrong frame or for which the two classes are not applicable (see (A.25d)). Also, assign the class OTHER if the verb is followed by a particle (i.e., if it is a phrasal verb).

save (NP1 V NP2 NP3) You will annotate three classes for the verb *save*: GET (GET), BILL (BILL), and OTHER (O). Examples are given as follows:

(A.26) a. Frank saved me a lot of money. BILL
b. If you buy a washing machine, it will save you the trouble. GET
c. I saved my bank account this week. O

- Use the class BILL when the direct object refers to money or money related concepts (see (A.26a)).
- Use the class GET in all other cases (see (A.26b)).
- Assign the class OTHER to tokens which have either the wrong frame or for which the two classes are not applicable (see (A.26c)). Also, assign the class OTHER if the verb is followed by a particle (i.e., if it is a phrasal verb).

shoot (NP1 V NP2 NP3) You will annotate three classes for the verb *shoot*: GET (GET), THROWING (THR), and OTHER (O). Examples are given as follows:

(A.27) a. Frank shot Mary a worried look. THR
b. I will shoot you some ducks. GET
c. We are shooting an episode every month. O

- Use the class GET when we are shooting something for someone (see (A.27b)).
- Use the class THROWING in all other cases (see (A.27a)).
- Assign the class OTHER to tokens which have either the wrong frame or for which the two classes are not applicable (see (A.27c)). Also, assign the class OTHER if the verb is followed by a particle (i.e., if it is a phrasal verb).

take (NP1 V NP2 NP3) You will annotate three classes for the verb *take*: BRING AND TAKE (BT), PERFORMANCE (PER), and OTHER (O). Examples are given as follows:

(A.28) a. John will take her the letter. BT
b. I took me weeks to finish the paper. PER
c. The government took several actions this year. O

- Use the class BRING AND TAKE when we take something to someone. In other words, when there is some movement involved (see (A.28a)).
- Use the class PERFORMANCE when we take something for someone, i.e., in their favor. Also, use the same class when something takes us time (see (A.28b)).
- Assign the class OTHER to tokens which have either the wrong frame or for which the two classes are not applicable (see (A.28c)). Also, assign the class OTHER if the verb is followed by a particle (i.e., if it is a phrasal verb).

produce (NP1 V NP2 *for* NP3) You will annotate three classes for the verb *produce*: CREATE (CR), PERFORMANCE (PER), and OTHER (O). Examples are given as follows:

(A.29) a. I produced an experiment for my students. CR
b. John produced a documentary for the BBC. PER
c. They produce money for more. O

- Assign the class CREATE when we are producing something for someone. When we are bringing into existence a new object for someone (see (A.29a)).
- Assign the class PERFORMANCE when we are producing something for someone within the context of a performance (see (A.29b)).
- Assign the class OTHER to tokens which have either the wrong frame or for which the two classes are not applicable (see (A.29c)). Also, assign the class OTHER if the verb is followed by a particle (i.e., if it is a phrasal verb).

take (NP1 V NP2 *for* NP3) You will annotate three classes for the verb *take*: STEAL (ST), PERFORMANCE (PER), and OTHER (O). Examples are given as follows:

(A.30) a. Take a sweater for Mary. PER
b. Frank has taken a course for his friend. ST
c. Jane has taken responsibility for her actions. O

- Use the PERFORMANCE class when the verb can participate in the benefactive alternation (e.g., *Take Mary a sweater*, see (A.30a)).
- Use the STEAL class when the verb cannot participate in the benefactive alternation (e.g., **Frank has taken his friend a course*, see (A.30b)).
- Assign the class OTHER to tokens which have either the wrong frame or for which the two classes are not applicable (see (A.30c)). Also, assign the class OTHER if the verb is followed by a particle (i.e., if it is a phrasal verb).

assess (NP1 V NP2) You will annotate three classes for the verb *assess*: ASSESSMENT (ASS), PRICE (PRI), and OTHER (O). Examples are given as follows:

(A.31) a. The police came to assess the situation. ASS
b. I would like to assess the cost. PRI
c. John assessed the extent. O

- Assign the class ASSESSMENT when the verb *assign* means "evaluate" (see (A.31a)).
- Assign the class PRICE when the object of the verb *assign* refers to cost- or value-related concepts (see (A.31b)).
- Assign the class OTHER when the meaning of the verb *assign* is underspecified (see (A.31c)) or when the noun phrase following the verb is not its object. Also, assign the class OTHER if the verb is followed by a particle (i.e., if it is a phrasal verb).

bang (NP1 V NP2) You will annotate three classes for the verb *bang*: HIT (HIT), SOUND EMISSION (SE), and OTHER (O). Examples are given as follows:

(A.32) a. Ethel banged her head on the wall. HIT
b. John banged the door as he left the room. SE
c. Christopher banged Rebecca. O
d. John banged something. O

- Assign the class HIT when the object of the verb *bang* is a body part (e.g., head, hand, etc.).
- Assign the class SOUND EMISSION when the object of the verb *bang* is an artefact (e.g., door, desk, glass, etc.).
- Assign the class OTHER when the verb *bang* has a meaning other than SOUND EMISSION or HIT (see (A.32c) where *bang* means "have sex with"). In general, assign the class OTHER when the object of the verb *bang* is animate.
- Assign the class OTHER when the meaning of the verb *bang* cannot be determined from its complement (see (A.32d)). Also, assign the class OTHER if the verb is followed by a particle (i.e., if it is a phrasal verb).

bite (NP1 V NP2) You will annotate three classes for the verb *bite*: HURT (HU), SWAT (SW), and OTHER (O). Examples are given as follows:

(A.33) a. Ellen bit her finger. HU
b. Cameron bit the pencil. SW
c. The baby is biting a lot. O

- Assign the class HURT when the object of the verb *bite* is a body part (see (A.33a)).
- Assign the class SWAT (which means "hit") when the object of the verb *bite* is an artefact (see (A.33b)).
- Assign the class other OTHER either when the meaning of *bite* cannot be determined from its complement or when the noun phrase following the verb *bite* is not its object (see (A.33c)). Also, assign the class OTHER if the verb is followed by a particle (i.e., if it is a phrasal verb).

clip (NP1 V NP2) You will annotate three classes for the verb *clip*: BRAID (BR), CUT (C), and OTHER (O). Examples are given as follows:

(A.34) a. I clipped my nails yesterday. BR
b. Monica clipped the tree branches yesterday. C
c. Peter clipped his shoes. O
d. John clipped a corner. O
e. Are you going to clip something? O

- Assign the class BRAID (which means "cut") when the object of the verb *clip* is a body part (e.g., nails, wings, ears, etc.) (see (A.34a)).
- Assign the class CUT when *clip* means "cut" and its object is not a body part (see (A.34b)).
- Assign the class OTHER when *clip* means "attach" (see (A.34c)).
- Assign the class OTHER when *clip* means "run over" (see (A.34d)).
- In general, assign the class OTHER in cases where *clip* does not mean "cut" and in cases where the meaning of *clip* is underspecified (see (A.34e)). Also, assign the class OTHER if the verb is followed by a particle (i.e., if it is a phrasal verb).

crack (NP1 V NP2) You will annotate three classes for the verb *crack*: BREAK (BR), SOUND EMISSION (SE), and OTHER (O). Examples are given as follows:

(A.35) a. Mary cracked the door. BR
b. Bill cracked the egg. BR
c. The government's policy managed to crack the union. BR
d. Patricia cracked the whip. SE

- You will assign the class BREAK when, as a result of cracking, the verb's object or parts of it are broken (see (A.35a–b)).

- You will Also, assign the class BREAK when the verb is used metaphorically (see (A.35c)).
- You will assign the class SOUND EMISSION when the object of the verb *crack* produces a noise without breaking (see (A.35d)).
- You will assign the class OTHER if the verb is followed by a particle (i.e., if it is a phrasal verb) and when the verb *crack* means:
  - "solve" (e.g., crack the problem, crack the code);
  - "joke" (e.g., crack jokes);
  - "open" (e.g., crack the bottle);
  - "break through" (e.g., crack barriers);
  - any other instance where *crack* means neither BREAK nor SOUND EMISSION.

crash (NP1 V NP2) You will annotate three classes for the verb *crash*: BREAK (BR), SOUND EMISSION (SE), and OTHER (O). Examples are given as follows:

(A.36)
a.	Rachel crashed the car.	BR
b.	The soldiers crashed the local resistance.	BR
c.	Peter crashed a shot.	SE
d.	Mary crashed the other day.	O
e.	Bob crashed something.	O

- You will assign the class BREAK when *crash* means "collide", "clash", or "break up" (see (A.36a)).
- You will Also, assign the class BREAK when *crash* is used metaphorically (see (A.36b)).
- You will assign the class SOUND EMISSION when the object of the verb *crash* produces a noise without breaking (see (A.36c)).
- You will assign the class OTHER when the verb *crash* is followed by the particle *down* (e.g., crash down) or the noun *land* (e.g., crash land).
- You will assign the class OTHER when the noun phrase following *crash* is not its object or when the meaning of the verb is underspecified (see (A.36e)).
- You will assign the class OTHER if the verb is followed by a particle (i.e., if it is a phrasal verb) and when the verb *crash* means:
  - "break through" (e.g., crash barriers);

- "intrude" (e.g., crash the party);
- "bring" (e.g., crash the ball home).

demolish (NP1 V NP2) You will annotate three classes for the verb *demolish*: AMUSE (AM), DESTROY (DE), and OTHER (O). Examples are given as follows:

(A.37) a. The priest's good words demolished her anger. AM
b. Ruth demolished the house. DE
c. We have demolished thousands. O

- Assign the class AMUSE only when the object of *demolish* is an abstract word (see (A.37a)). In general assign the class AMUSE when the object of *demolish* cannot be physically destroyed.
- Assign the class DESTROY when the object of *demolish* is an artefact (see (A.37b)) or generally a thing which can be physically destroyed.
- Assign the class OTHER when the noun phrase following *demolish* is not its object or when the meaning of *demolish* is underspecified (see (A.37c)). Also assign the class OTHER if the verb is followed by a particle (i.e., if it is a phrasal verb).

draw (NP1 V NP2) You will annotated three classes for the verb *draw*: SCRIBBLE (SCR), PULL (PUL), and OTHER (O). Examples are given as follows:

(A.38) a. The child drew her mother's dress under the table. PUL
b. Frank wants to draw attention. PUL
c. He will draw a number. PUL
d. Kate likes to draw dogs. SCR
e. I will draw everything. O

- Assign the class PULL, when the verb *draw* means "pull" (see (A.38a)), "attract" (see (A.38b)), or "select" (see (A.38c)).
- Assign the class SCRIBBLE when *draw* means "make a drawing" (see (A.38d)).
- Assign the class OTHER when the meaning of the verb *draw* is either underspecified (see (A.38e)) or when the noun phrase following *demolish* is not its object. Also assign the class OTHER if the verb is followed by a particle (i.e., if it is a phrasal verb).

feel (NP1 V NP2) You will annotate three classes for the verb *feel*: HUNT (HU), SEE (SEE), and OTHER (O). Examples are given as follows:

(A.39) a. The doctor felt my tummy for lumps. HU
b. Why do I have to feel so much pain? SEE
c. Mary feels peaceful now. O
d. Peter feels that way. O
e. I can feel something. O

- You will assign the class SEE when the verb *feel* means "experience" and is followed by a noun which is its object (see (A.39b)).
- You will assign the class HUNT when the verb *feel* means "look for" or "touch" (see (A.39a)).
- You will assign the class OTHER if: (a) the verb *feel* is predicative (see (A.39c)), (b) the token has a wrong frame (see (A.39d)), (c) the meaning of the verb is underspecified (see (A.39e)), and (d) the verb is followed by a particle (i.e., it is a phrasal verb).

file (NP1 V NP2) You will annotate four classes for the verb *file*: BRAID (BR), CARVE (CAR), POCKET (POC), and OTHER (O). Examples are given as follows:

(A.40) a. Sarah is filing her nails at the moment. BR
b. Peter is filing the metal. CAR
c. I will file this year's expenses. POC
d. Carry filed the charges yesterday. O

- You will assign the class BRAID when the object of the verb *file* is a body part (e.g., nails, see (A.40a)).
- You will assign the class CARVE when the verb *file* means "smoothen" (see (A.40b)).
- You will assign the class POCKET when the verb *file* means "register" (e.g., file a complaint, file a record, file charges, see (A.40c)).
- You will assign the class OTHER if: (a) the verb *file* has the meaning "initiate legal action" (see (A.40d)), (b) the meaning of the verb is underspecified, (c) the noun phrasing following *file* is not its object, and (d) the verb is followed by a particle (i.e., it is a phrasal verb).

grind (NP1 V NP2) You will annotate four classes for the verb *grind*: BUILD (BUI), CARVE (CAR), CRANE (CR), and OTHER (O). Examples are given as follows:

(A.41) a. John will grind the corn. BUI
b. You have to grind this knife. CAR
c. Pamela is grinding her teeth when she sleeps. CR

d. My husband will grind his sister for what she did. O
e. Roger tried to grind his anger. O
f. Grind your cigarette out. O
g. Tim is grinding his words. O

- You will assign the class BUILD when *grind* means "fragment" and its object is neither an animate nor an abstract noun (see (A.41a)).
- You will assign the class CARVE when *grind* means "sharpen" (see (A.41b)).
- You will assign the class CRANE when the object of the verb *grind* is a body part (see (A.41c)).
- You will assign the class OTHER if: (a) the object of *grind* is an abstract or animate noun. (see (A.41d,e)), (b) the noun phrase following the verb is not its object, (c) the meaning of the verb is "rub" (see (A.41f)), (d) neither of the three classes (CRANE, CARVE, BUILD) can be assigned (see (A.41g)), and (e) the verb is followed by a particle (i.e., it is a phrasal verb).

insult (NP1 V NP2) You will annotate three classes for the verb *insult*: AMUSE (AM), JUDGMENT (JUD), and OTHER (O). Examples are given as follows:

(A.42) a. Max insulted his daughter. AM
b. Jane insulted the constitution. JUD
c. He insulted everything. O

- Assign the class AMUSE when the object of the verb *insult* is animate (see (A.42a)).
- Assign the class JUDGMENT when the object of *insult* is not animate (see (A.42b)).
- Assign the class OTHER when the noun phrase following the verb is not its object or when the meaning of the verb is underspecified (see (A.42c)). Also, assign the class OTHER if the verb is followed by a particle (i.e., if it is a phrasal verb).

kick (NP1 V NP2) You will annotate three classes for the verb *kick*: HIT (HIT), THROW (THR), and OTHER (O). Examples are given as follows:

(A.43) a. John kicked the table. HIT
b. Mary kicked the ball. THR
c. Jeremy kicks arse. O

- Assign the class HIT when the object of the verb *hit* is either animate or a body part, or an artefact which cannot be easily thrown (see (A.43a)). For example, we don't throw tables to each other.

- Assign the class THROW when the object of the verb *kick* is an artefact which is thrown by being kicked. For instance, we kick balls, tins, goals (see (A.43b)).
- Assign the class OTHER when the meaning of *kick* is neither HIT nor THROW (see (A.43c)) or when the noun phrase following *kick* is not its object. Also, assign the class OTHER if the verb is followed by a particle (i.e., if it is a phrasal verb).

lick (NP1 V NP2) You will annotate three classes for the verb *lick*: TOUCH (TOU), WIPE MANNER (WM), and OTHER (O). Examples are given as follows:

(A.44)	a.	Catherine licks my hands.	TOU
	b.	Catherine licked Bob.	TOU
	c.	Mary licked her ice-cream.	WM
	d.	Frank licked the bowl.	WM
	e.	Bill licked the floor.	WM
	f.	Ellen licks her boss.	O

- Assign the class TOUCH when the object of the verb *lick* is a body part (see (A.44a)) or animate (see (A.44b)).
- Assign the class WIPE MANNER when the object of *lick* is food (see (A.44c)), or a food container or a food-related object (see (A.44d)).
- Assign the class WIPE MANNER when the act of licking suggests the removal of a substance from the object being licked (see (A.44c)–(A.44e)).
- Assign the class OTHER if: (a) the verb *lick* means neither TOUCH nor WIPE MANNER (see (A.44f)), (b) the noun phrase following *lick* is not its object, and (c) the verb is followed by a particle (i.e., it is a phrasal verb).

miss (NP1 V NP2) You will annotate three classes for the verb *miss*: ADMIRE (ADM), CONTIGUOUS LOCATION (CL), and OTHER (OTH). Examples are given as follows:

(A.45)	a.	Rachel misses her sister.	ADM
	b.	Tom missed the bus.	CL
	c.	Helen missed her lesson.	CL
	d.	I missed the other day.	O

- You will assign the class ADMIRE when the verb *miss* means "desire" or "feel the absence of" (see (A.45a)).
- You will assign the class CONTIGUOUS LOCATION when *miss* means "lose", "fail to reach", or "fail to get" (see (A.45b,c)).

- You will assign the class OTHER when the noun phrase following the verb *miss* is not its object (see (A.45d)). Also, assign the class OTHER if the verb is followed by a particle (i.e., if it is a phrasal verb).

move (NP1 V NP2) You will annotate three classes for the verb *move*: ADMIRE (ADM), ROLL (RO), SLIDE (SL), and OTHER (O). Examples are given as follows:

(A.46) a. John moved his sister in law. ADM
b. Martha moved her eyes. RO
c. I moved my clothes. SL
d. Could you please move your head? SL
e. We moved five miles O

- You will assign the class ADMIRE when *move* means "affect" or "stir in emotions" (see (A.46a)).
- Assign the class ROLL only when the object of the verb *move* can naturally roll (see (A.46b)).
- Assign the class SLIDE in all other cases (i.e., when the object of *move* is moved, see (A.46c,d)).
- Assign the class OTHER when the noun phrase following *move* is not its object (see (A.46e)). Also, assign the class OTHER if the verb is followed by a particle (i.e., if it is a phrasal verb).

poke (NP1 V NP2) You will annotate three classes for the verb *poke*: POKE (POK), RUMMAGE (RUM), and OTHER (O).

(A.47) a. The doctor poked her belly for lumps. RUM
b. Bill pokes his children all the time. POK
c. The neighbours poked their heads out of the windows. POK
d. He is always poking something. O
e. I will go poke my way. O

- You will assign the class RUMMAGE when the verb *poke* means "look for" (see (A.47a)).
- You will assign the class POKE in all other cases. For example when *poke* means "hit" (see (A.47b)), or "stick out" (see (A.47c)).
- You will assign the class OTHER if: (a) the meaning of *poke* is underspecified (see (A.47d)), (b) the noun phrase following *poke* is not its object (see (A.47e)), and (c) the verb is followed by a particle (i.e., it is a phrasal verb).

push (NP1 V NP2) You will annotate three classes for the verb *push*: CARRY (CAR), PUSH PULL (PP), and OTHER. Examples are given as follows:

(A.48) a. I pushed the bed with Mary. CAR
b. Max pushed Rachel. PP
c. John pushed the bell. PP
d. Let me push my way. O

- You will assign the class CARRY only when the verb *push* is followed by the preposition *with* (see (A.48a)).
- In all other cases you will assign the class PUSH PULL (see (A.48b,c)).
- Assign the class OTHER when: (a) the meaning of *push* is neither PUSH PULL nor CARRY, (b) when the noun phrase following *push* is not its object (see (A.48d)), and (c) the verb is followed by a particle (i.e., it is a phrasal verb).

rub (NP1 V NP2) You will annotate three classes for the verb *rub*: CRANE, WIPE MANNER, and OTHER. Examples are given as follows:

(A.49) a. Mary rubbed her hands and face. CR
b. Peter rubbed the donkey. WM
c. The fire rubbed out the entire village. O

- You will assign the class CRANE when the object of the verb *rub* is a body part (see (A.49a)).
- You will assign the class WIPE MANNER when the object of *rub* is not a body part (see (A.49b)).
- You will assign the class OTHER if *rub* has neither of the above two meanings and Also, if it is followed by a particle (i.e., if it is a phrasal verb, see (A.49c)).

salute (NP1 V NP2) You will annotate three classes for the verb *salute*: CURTSEY (CUR), JUDGMENT (JUD), and OTHER (O). Examples are given as follows:

(A.50) a. Max saluted the dog. CUR
b. The president saluted Veronica's achievements. JUD
c. Catherine will salute next month. O

- You will assign the class CURTSEY when the object of the verb *salute* is animate (see (A.50a)).

- You will assign the class JUDGMENT when the object of *salute* is not animate (see (A.50b)).
- You will assign the class OTHER when the noun phrase following *salute* is not its object (see (A.50c)) or when the meaning of *salute* is neither CURTSEY nor JUDGMENT. Also, assign the class OTHER if the verb is followed by a particle (i.e., if it is a phrasal verb).

savour (NP1 V NP2) You will annotate three classes for the verb *savour*: ADMIRE (ADM), SIGHT (SIG), and OTHER. Examples are given as follows:

(A.51) a. John savoured the landscape. SIG
b. Rachel savoured the cake. ADM
c. I savour every two minutes. O

- You will assign the class SIGHT only when we admire something through seeing (see (A.51a)).
- In all other cases (i.e., when *savour* means "taste" or "smell"), you will assign the class ADMIRE (see (A.51b)).
- You will assign the class OTHER when the noun phrase following *salute* is not its object (see (A.51c)) or when the meaning of *salute* is neither SIGHT nor ADMIRE. Also, assign the class OTHER if the verb is followed by a particle (i.e., if it is a phrasal verb).

scrape (NP1 V NP2) You will annotate three classes for the verb *scrape*: CUT (C), WIPE MANNER (WM), and OTHER. Examples are given as follows:

(A.52) a. Will you let me scrape your hand? C
b. Elisabeth will scrape the bucket. WM
c. I can scrape anything. O
d. We want to scrape acquaintance. O

- You will assign the class CUT when the object of the verb *scrape* is a body part (see (A.52a)).
- You will assign the class WIPE MANNER when the object of *scrape* is not a body part (see (A.52b)).
- You will assign the class OTHER when: (a) the meaning of *scrape* is underspecified (see (A.52c), (b) the verb *scrape* means neither CUT nor WIPE MANNER (see (A.52d)), and (c) the verb is followed by a particle (i.e., it is a phrasal verb).

scratch (NP1 V NP2) You will annotate three classes for the verb *scratch*: HURT (HU), WIPE MANNER (WM), and OTHER. Examples are given as follows:

(A.53) a. The doctor scratched his beard. HU
b. John scratched his sister. HU
c. I can scratch the surface. WM
d. We want to scratch something. O

- You will assign the class HURT when the object of the verb *scratch* is a body part (see (A.53a)) or animate (see (A.53b)).
- You will assign the class WIPE MANNER when the object of *scratch* is neither a body part nor animate (see (A.53c)).
- You will assign the class OTHER if: (a) the meaning of *scratch* is underspecified (see (A.53d), (b) the verb *scratch* means neither HURT nor WIPE MANNER, and (c) if the verb is followed by a particle (i.e., if it is a phrasal verb).

scrutinize (NP1 V NP2) You will annotate three classes for the verb *scrutinize*: ASSESSMENT (ASS), SIGHT (SIG), and OTHER. Examples are given as follows:

(A.54) a. The lawyers scrutinized the case. ASS
b. I scrutinized the book. ASS
c. Barbara scrutinized the baby. SIG
d. Mary scrutinized the body. SIG
e. Bill scrutinized the rest. O

- You will assign the class ASSESSMENT when the verb *scrutinize* means "examine", "analyze", or "study" (see (A.54a,b)). Usually the object of *scrutinize* is an abstract noun.
- You will assign the class SIGHT when the object of the verb *scrutinize* is either an artefact or animate. In general, assign the class SIGHT for things we can see (see (A.54c,d)).
- You will assign the class OTHER when the meaning of *scrutinize* is underspecified (see (A.54e)). You will Also, assign the class OTHER if the verb is followed by a particle (i.e., if it is a phrasal verb).

shoot (NP1 V NP2) You will annotate three classes for the verb *shoot*: POISON (POI), SWAT (SW), THROW (THR), and OTHER (O). Examples are given as follows:

(A.55) a. Pamela shot Frank and the dog. POI
b. Rebecca shot the orange. SW
c. Pam likes shooting darts. THR
d. I like shooting pool. O
e. We shot a film yesterday. O

- Assign the class POISON when we shoot to kill someone. Usually the object of *shoot* is animate (see (A.55a)).
- Assign the class SWAT when *shoot* means shoot a target (see (A.55b)). Usually the object of *shoot* is not animate.
- Assign the class THROW when the object of *throw* is being ejected (e.g., shoot an arrow, shoot a marble, see (A.55c)).
- Assign the class OTHER if: (a) the meaning of *shoot* is neither POISON nor THROW nor SWAT (see (A.55d,e)) and (b) the verb is followed by a particle (i.e., it is a phrasal verb).

stab (NP1 V NP2) You will annotate three classes for the verb *stab*: HIT (HIT), POISON (POI), and OTHER (O). Examples are given as follows:

(A.56) a. John stabbed the desk. HIT
b. Barbara stabbed her finger HIT
c. Martin stabbed Scott and his dog. POI
d. He will stab every month. O

- You will assign the class HIT when the object of *stab* is either an artefact (see (A.56a)) or a body part (see (A.56b)).
- You will assign the class POISON when the object of the verb *stab* is animate (see (A.56c)).
- You will assign the class OTHER when the noun phrase following *stab* is not its object. You will Also, assign the class OTHER if the verb is followed by a particle (i.e., if it is a phrasal verb).

smash (NP1 V NP2) You will annotate three classes for the verb *smash*: BREAK (BR), HIT (HIT), and OTHER. Examples are given as follows:

(A.57) a. My father smashed the piano. BR
b. His complaints smashed my patience. BR
c. John smashed his head. HIT
d. Betty smashed the team. HIT
e. Let me smash something. O

- Assign the class BREAK when the object of *smash* is an artefact (see (A.57a)). Assign the class BREAK also when *smash* is used metaphorically (see (A.57b)).
- You will assign the class HIT when the object of *smash* is either a body part (see (A.57c)) or animate (see (A.57d)).
- Assign the class BREAK if: (a) *smash* means neither BREAK nor HIT, (b) the meaning of the verb *smash* is underspecified (see (A.57e)), (c) the noun phrase following *smash* is not its object, and (d) the verb is followed by a particle (i.e., it is a phrasal verb).

study (NP1 V NP2) You will annotate four classes for the verb *study*: ASSESSMENT (ASS), LEARN (LEA), SIGHT (SIG), and OTHER. Examples are given as follows:

(A.58) a. We studied the report carefully. ASS
b. Bob studies astrophysics. LEA
c. Mary studied her face in the mirror. SIG
d. Sue studied three times. O

- Assign the class ASSESSMENT when the verb *study* means "examine" or "analyze" (see (A.58a)).
- Assign the class LEARN when we study a discipline or take a course (see (A.58b)).
- Assign the class SIGHT only when we study something by inspection, by "seeing" it (see (A.58c)).
- Assign the class OTHER when the noun phrase following *study* is not its object (see (A.58d)). Also, assign the class OTHER if the verb is followed by a particle (i.e., if it is a phrasal verb).

suck (NP1 V NP2) You will annotate three classes for the verb *suck*: CHEW (CH), WIPE MANNER (WM), and OTHER (O). Examples are given as follows:

(A.59) a. Billy still sucks his thumb. CH
b. Margaret sucked the sun on the Spanish islands. WM
c. He always sucks something. O

- Assign the class CHEW when the object of the verb *suck* can be chewed (see (A.59a)).
- Assign the class WIPE MANNER when the object of *suck* can be drank or absorbed (see (A.59b)).
- Assign the class OTHER for the expression "suck someone's teeth". Also, assign the class OTHER when the meaning of the verb is underspecified (see (A.59c)), when the noun phrase following *suck* is not its object or when the verb is followed by a particle (i.e., it is a phrasal verb).

support (NP1 V NP2) You will annotate three classes for the verb *support*: ADMIRE (ADM), CONTIGUOUS LOCATION (CL), and OTHER (O). Examples are given as follows:

(A.60) a. I support the president of Africa. ADM
b. Mike supports the decision. ADM
c. The table will not support your weight. CL
d. Everyone supports something. O

- Assign the class ADMIRE when the object of the verb *support* is animate (see (A.60a)). Assign the class ADMIRE when support means "back" or "approve" (see (A.60b)).
- Assign the class CONTIGUOUS LOCATION when *support* means "sustain" or "hold up" (see (A.60c)).
- Assign the class OTHER when the meaning of *support* is underspecified (see (A.60d)). Also, assign the class OTHER if the verb is followed by a particle (i.e., if it is a phrasal verb).

toast (NP1 V NP2) You will annotate three classes for the verb *toast*: COOK (CK), JUDGMENT (JUD), and OTHER (O). Examples are given as follows:

(A.61) a. Catherine toasted the bread. CK
b. I toasted my back on the beach yesterday. CK
c. Let's toast your promotion. JUD
d. The employees toasted Jonathan. JUD
e. I will toast everything. O

- Assign the class COOK when the object of the verb *toast* is related to food (see (A.61a)) or when the verb has the meaning "burn". In the latter case the objects of *cook* are usually body parts (see (A.61b)).
- Assign the class JUDGMENT when the verb *toast* means "drink to" or "honor" (see (A.61c–d)).
- Assign the class OTHER when the meaning of *toast* is neither COOK nor JUDGMENT or when the meaning of the verb is underspecified (see (A.61e)). Also, assign the class OTHER if the verb is followed by a particle (i.e., if it is a phrasal verb).

tug (NP1 V NP2) You will annotate three classes for the verb *tug*:CARRY (CAR), PUSH PULL (PP), and OTHER (O). Examples are given as follows:

(A.62) a. Patrick tugged his bike home. CAR
b. Frank tugged his coat. PP

c. Mary tugged Betty. PP
d. Bob tugged his leg free. PP
e. My fingers tugged down Bill's body. O

- You will assign the class CARRY when the verb *tug* means "transport" or "carry" (see (A.62a)).
- You will assign the class PUSH PULL when the verb *tug* means "draw" or "pull" (see ((A.62b)–(A.62d)).
- You will assign the class OTHER when *tug* means neither PUSH PULL nor CARRY (see (A.62e)). You will Also, assign the class OTHER if the verb is followed by a particle (i.e., if it is a phrasal verb).

value (NP1 V NP2) You will annotate three classes for the verb *value*: ADMIRE (ADM), PRICE (PRI), and OTHER (O). Examples are given as follows:

(A.63) a. Bob values Mary's judgment. ADM
b. The president valued the company, and its assets at 6 billion pounds. PRI
c. Rachel will value something. O

- Assign the class ADMIRE when *value* means "appreciate" or "treasure" (see (A.63a)).
- Assign the class PRICE when the verb *value* means "assess" or "evaluate". Usually it is the value of the object of the verb which is assessed (see (A.63b)).
- Assign the class OTHER either when the noun phrase following *value* is not its object or when the meaning of the verb is underspecified (see (A.63c)). Also, assign the class OTHER if the verb is followed by a particle (i.e., if it is a phrasal verb).

whisk (NP1 V NP2) You will annotate three classes for the verb *whisk*: SPANK (SP), WIPE MANNER (WM), and OTHER (O). Examples are given as follows:

(A.64) a. Elisabeth whisked the president. SP
b. I whisked the egg-whites. WM
c. Let me whisk something. O

- Assign the class SPANK when the object of the verb *whisk* is animate (see (A.64a)).
- Assign the class WIPE MANNER when the object of *whisk* is not animate (see (A.64b)).
- Assign the class OTHER either when *whisk* means neither SPANK nor WIPE MANNER or when its meaning is underspecified (see (A.64c)). Also, assign the class OTHER if the verb is followed by a particle (i.e., if it is a phrasal verb).

kick (NP1 V *at* NP2) You will annotate three classes for the verb *kick*: BODY INTERNAL MOTION (BIM), HIT (HIR), and OTHER (O). Examples are given as follows:

(A.65) a. Mary kicked at the start. BIM
b. Barbara kicked at the table. HIT
c. The riot kicked off at the president's office. O

- Assign the class BODY INTERNAL MOTION when the PP following the verb is locative or temporal (see (A.65a)).
- Assign the class HIT when the PP following the verb is directive (see (A.65b)).
- Assign the class OTHER when the meaning of the verb is underspecified. Also, assign the class OTHER if the verb is followed by a particle (i.e., if it is a phrasal verb, see (A.65c)).

A.2. Acquisition of Compound Nouns

You will be given a series of noun-noun sequences as shown in example (A.66). The noun-noun sequence *plane cabin* is given in the sentence in which it occurs. In addition, you will be provided with the sentence preceding and following the sentence in which the sequence is found. The noun-noun sequence will be always within brackets.

(A.66) There's staff with uniforms to look after you. A [plane cabin] tries to fool you with the same set-up, but suddenly it meets turbulence, bumps and jolts, and three hundred of you sit there thinking of the drop beneath. In the middle of A Man for All Seasons there's a ping, and on come the little red signs: FASTEN YOUR SEATBELT.

You will annotate the noun-noun sequence with two classes: C (for sequences that are valid compounds) and N (for sequences that are not compounds). Here are some heuristics to help you decide whether the sequence is a compound or not:

- **N** if the noun-noun sequence is the result of tagging mistake. This means that one of the two words in the sequence has been mistagged as a noun instead of a verb (see (A.67a)), a gerund (see (A.67b)), or an adjective (see (A.67c)).
- **N** if one of the nouns in the sequence is a proper name (see (A.68a)) or an acronym (see (A.68b)).
- **N** if one or both words in the sequence are foreign (i.e., non-English) terms (see (A.68c)).
- **N** if the noun-noun sequence is the result of a parsing (bracketing) error, i.e., the two nouns do not form a constituent (see the examples in (A.69)).

- **N** if the noun-noun sequence is preceded by an adjective or noun modifying the left-most noun in the sequence instead of the rightmost one (see the examples in (A.70))
- **C** in all other cases.

(A.67) a. By contrast Krystyna Ziach's three photographically-based [installations delight] primarily through their formal qualities.
b. He had felt them—round [swelling lumps] the size of gold coins.
c. In my own work with [Japanese quail] I have found that this process may lead to a preference for a partner that is slightly novel—just a bit different but not too different from the members of the opposite sex it knew when it was young (Bateson, Mate Choice).

(A.68) a. The Italian government, which takes over the EC presidency from Ireland next summer, will use the three-stage [Delors strategy] for monetary union as the basis for considering changes to the Treaty of Rome.
b. The response of the [CNAA officer] who visited Brighton was frankly incredulous.
c. I am not able to acquit the Scots of this fault (pride), he wrote; and "ill est [fier comme] ung Escossoys" (he is as proud as a Scot) was his record of what the French thought of the matter.

(A.69) a. Susan and Luke and I were [friends years] ago.
b. Consequently in the first six [years implementation] was patchy.

(A.70) a. God I need a time and motion expert in this cupboard. Well it's multi [purpose cupboard].
b. No matter where in the world human [rights violations] occur – from India to Iran, Chile to Czechoslovakia.
c. Any programme of final drama [school productions] will present a variety of styles.

Note: It may help you decide whether the two nouns form compound or not if you have a look at their parts of speech.

A.3. Interpretation of Nominalizations

You will be given a series of nominalizations (compound nouns whose heads are derived from verbs) as shown in example (A.71). The compound *data holder* is given in the sentence in which it occurs. In addition, you will be provided with the sentence which precedes and follows the sentence in which the compound is found. The nominalizations will be always within brackets.

(A.71) To find out four people all well known in their fields agreed to let us find the details of their financial, health, police, and other records. The results of our investigation

are deeply disturbing, no individual in the country can be sure that their secrets are safe, banks and other [data holders] seem powerless to stop the growth in this trade and the Government is unwilling to crack down on it. We started our investigation by asking Robert to come and see us, we prepared for his visit by hiding a camera in this umbrella, which lay on top of a brief case.

You will annotate the compound with two classes: subject (SUBJ) and object (OBJ). For each compound do the following:

- Convert the rightmost noun of the compound into a verb. Suppose the compound is *data holder*, then convert *holder* into the verb *hold*.
- Try to determine whether the noun modifying the compound head is the subject or the object of the verb in question.

In the case of *data holder* try to determine if *data* is the subject or the object of the verb *hold*. It is helpful to ask questions like the following:

- Is someone/something holding the data?
- Is the data holding something/someone?

Given this particular example, it is more natural to assume that it is someone/something holding the data rather than it is the data which are holding someone/something. In fact, if you read carefully the text in A.71, you understand that it is the bank which is holing the data.

Consider now the compound *employment behaviour* given in example (A.72). Here, it is easy to figure out that *employment* is the subject of the verb *behave*, simply because *behave* is an intransitive verb. In other cases it is not so straightforward. Consider the nominalization *industry reception* in (A.73). Questions like the following seem equally plausible:

- Is someone/something receiving the industry?
- Is the industry receiving someone/something?

It is only after you read the full context that you realize that it is the industry who is having the reception.

(A.72) A variety of statistical procedures will be explored, which are appropriate to the analysis of such longitudinal data and which overcome this problem of omitted factors, amongst others. Analyses will be undertaken using these methods to identify some of the important factors determining migration and [employment behaviour] in the British Isles. Postdoctor Research Fellowships.

(A.73) This trip included a review of the company's community aid programme by public relations manager Maria Teresa de Angel, as well as the presentation of long service awards to 12 members of LASMO 's field staff (see story on page 11). The final evening saw more than 300 guests attend an [industry reception], hosted by LASMO. Commenting on the tour, Graeme Stephens said: This visit was a huge success.

In sum, here are some heuristics that will help you decide whether the modifier is the subject or object of the compound head:

1. convert the nominalized head into a verb;
2. ask questions of the sort "who is doing what?"
3. usually, if the head ends in *-er, -or, -ant*, the modifier is the object; if the heads ends in *-ee*, the modifier is the subject;
4. if the verb is intransitive the modifier is the subject;
5. if you can't decide read carefully the context and use common knowledge;
6. finally, if you are still undecided, default to object.

Appendix B

Instructions

In what follows we give the instructions for Experiment 8 (see Chapter 5).

Thanks for taking part in this experiment!

Please read the instructions carefully before starting. Do not hesitate to contact the experimenter in case you have any questions or comments concerning this experiment.

This experiment requires a Java compatible web browser and Java has to be enabled.

Depending on the hardware and browser you use, and on your net connection, the execution of the experiment may be slow at times.

Personal Details

As part of this experiment, we have to collect a small amount of personal information, which we ask you to enter in the Personal Details window below. *This information will be treated confidential, and will not be made available to a third party. None of the responses collected in this experiment will be associated with your name in any way.* If you have any questions about this practice, please contact the experimenter.

Please be careful to fill in the Personal Details questionnaire correctly, as otherwise we will have to discard your responses.

We ask you to supply the following information:

- your name and email address;
- your age and sex;

- whether you are right or left handed (based on the hand you prefer to use for writing);
- the academic subject you study or have studied (or your current occupation in case you haven't attended university);
- under 'Region', please specify the place (city, region/state/province, country) where you have learned your first language.

Instructions

Part 1: Judging Line Length

Before doing the main part of the experiment, you will do a short task involving judging line length. A series of lines of different length will be presented on the screen. Your task is to estimate how long they seem by assigning numbers to them. You are supposed to make your estimates relative to the first line you will see, your *reference* line. Give it any number that seems appropriate to you, bearing in mind that some of the lines will be longer than the reference and some will be shorter.

After you have judged the reference line, assign a number to each following line so that it represents how long the line is in proportion to the reference. The longer it is compared to the reference, the larger the number you will use; the shorter it is compared to the reference, the smaller the number you will use. So if you feel that a line is twice as long as the reference, give it a number twice the reference number; if it's a third as long, provide a number a third as big as the reference.

So, if the reference is this line, you might give it the number 10:

If you have to judge this line, you might assign it 17:

And this one might be 2.5:

There is no limit to the range of numbers you may use. You may use whole numbers or decimals. If you assigned the reference line the number 1, you might want to call the last one 0.25. Just try to make each number match the length of the line as you see it.

Parts 2 and 3: Judging Adjective-Noun Paraphrases

In the main part of the experiment, you will be asked to use numbers to judge how well a given sentence paraphrases the meaning of an adjective-noun combination. You will see a series of adjective-noun pairs presented with their paraphrases one at a time in the Experiment window below. Some paraphrases will seem perfectly OK to you, but others will not.

Your task is to judge how well the sentence paraphrases the adjective-noun combination by assigning a number to it. First you will see a *reference* paraphrase, to which you have to assign a reference number. You can use any number that seems appropriate to you. For each subsequent paraphrase of an adjective-noun pair, please assign a number to show how good or bad that paraphrase is in proportion to the reference.

For example, if the reference paraphrase was:

EASY CAR — a car which is difficult to meet

you would probably give it a rather low number. (You are free to decide what 'low' or 'high' means in this context.) If the next example:

EASY CAR — a car which is easy to drive

seemed 10 times better than the reference, you'd give it a number 10 times the number you gave to the reference. If it seemed half as good as the reference, you'd give it a number half the number you gave to the reference.

You can use any range of positive numbers that you like, including decimal numbers. There is no upper or lower limit to the numbers you can use, except that you cannot use zero or negative numbers. Try to use a wide range of numbers and to distinguish as many degrees of 'plausibility' of adjective-noun paraphrases as possible.

There are no 'correct' answers, so whatever seems right to you is a valid response. We are interested in your first impressions, so please don't take too much time to think about any one adjective-noun paraphrase: try to make up your mind quickly, spending less than 5 seconds on each paraphrase.

Procedure

First please fill in the Personal Details questionnaire as explained above, and then press the Start button.

The experiment will consist of the following 3 parts:

- Training session: judging 6 line lengths
- Practice session: judging 8 adjective-noun paraphrases
- Experiment session: judging 135 adjective-noun paraphrases

In each part you will see the reference item in the experiment window. Please enter your reference number and then press the Continue button. Now the test items will appear one after the other in the experiment window. Please type your judgment in the box below each item.

The experiment will take about 20 minutes. After the experiment is completed you will receive an email confirmation of your participation.

Please keep in mind:

- Use any number you like for the reference paraphrase.
- Judge each adjective-noun paraphrase in proportion to the reference.
- Use any positive numbers which you think are appropriate.
- Use high numbers for 'good' adjective-noun paraphrases, low numbers for 'bad' paraphrases and intermediate numbers for paraphrases which are intermediate in plausibility.
- Try to use a wide range of numbers and to distinguish as many degrees of paraphrase 'plausibility' as possible.
- Try to make up your mind quickly, base your judgments on your first impressions.

Appendix C

Materials

Examples for the stimuli used in Experiment 8 are given below. The a, b, and c examples correspond to High, Medium and Low interpretations. Tables C.1 and C.2 give the entire list of randomly selected verbs in the three experimental conditions (High, Medium, Low) together with elicited interpretation preferences. The elicited ratings are normalized and log-transformed.

Object interpretation

(C.1) a. difficult customer — a customer who is difficult to satisfy
b. difficult customer — a customer who is difficult to help
c. difficult customer — a customer who is difficult to drive

(C.2) a. easy task — a task that is easy to perform
b. easy task — a task that is easy to manage
c. easy task — a task that is easy to begin

(C.3) a. good language — a language that is good to know
b. good language — a language that is good to reinforce
c. good language — a language that is good to encourage

(C.4) a. hard logic — logic that is hard to understand
b. hard logic — logic that is hard to express
c. hard logic — logic that is hard to impose

(C.5) a. slow minute — a minute that one takes slowly
b. slow minute — a minute that one fills slowly
c. slow minute — a minute that one meets slowly

Subject interpretation

(C.6) a. difficult passage — a passage that reads with difficulty
b. difficult passage — a passage that speaks with difficulty
c. difficult passage — a passage that appears with difficulty

(C.7) a. easy car a car that starts easily
b. easy car a car that moves easily
c. easy car a car that closes easily

(C.8) a. fast horse a horse that runs fast
b. fast horse a horse that works fast
c. fast horse a horse that sees quickly

(C.9) a. good light a light that works well
b. good light a light that spreads well
c. good light a light that increases well

(C.10) a. slow child a child that reacts slowly
b. slow child a child that adapts slowly
c. slow child a child that expresses something slowly

The experimental item in (C.11) was presented as the modulus. The verb *alter* was selected from the interpretations which were derived by the model for the adjective-noun combination *hard substance* and corresponded to the Medium Probability band.

(C.11) hard substance a substance that is hard to alter

Adjective	Noun	High		Medium		Low	
difficult	consequence	cope with	.3834	analyse	.0270	refer to	−.3444
difficult	customer	satisfy	.4854	help	.3228	drive	−.4932
difficult	friend	live with	.2291	approach	.0798	miss	−.5572
difficult	group	work with	.3066	teach	.2081	respond to	.1097
difficult	hour	endure	.3387	complete	−.1386	enjoy	.1600
easy	comparison	make	.4041	discuss	−.0901	come to	.3670
easy	food	cook	.2375	introduce	−.3673	finish	−.1052
easy	habit	get into	.2592	explain	−.2877	support	−.0523
easy	point	score	.3255	answer	.1198	know	−.0307
easy	task	perform	.4154	manage	.3094	begin	.0455
fast	device	drive	−.0908	make	−.2948	see	−.4817
fast	launch	stop	−.5438	make	.0075	see	−.3963
fast	pig	catch	−.5596	stop	−.6285	use	−.5350
fast	rhythm	beat	.0736	feel	−.1911	make	−.1296
fast	town	protect	−.5896	make	−.4564	use	−.4996
good	climate	grow up in	.2343	play in	−.6498	experience	−.4842
good	documentation	use	.2549	produce	−.2374	include	.1110
good	garment	wear	.2343	draw	−.6498	measure	−.4842
good	language	know	.2188	reinforce	−.1383	encourage	−.0418
good	postcard	send	.1540	draw	−.3248	look at	.2677
hard	logic	understand	.2508	express	.0980	impose	−.2398
hard	number	remember	.0326	use	−.3428	create	−.4122
hard	path	walk	.2414	maintain	.0343	explore	.1830
hard	problem	solve	.4683	express	.0257	admit	−.2913
hard	war	fight	.2380	get through	.2968	enjoy	−.5381
slow	child	adopt	−.5028	find	−.7045	forget	−.6153
slow	hand	grasp	−.5082	win	−.2524	produce	−.3360
slow	meal	provide	−.0540	begin	−.2546	bring	−.3965
slow	minute	take	−.1396	fill	−.1131	meet	−.6083
slow	progress	make	.3617	bring	−.1519	give	−.2700
safe	building	use	.1436	arrive at	−.1640	come in	.0306
safe	drug	release	.1503	try	.1930	start	.1614
safe	house	go to	.2139	get	−.3438	make	−.3490
safe	speed	arrive at	−.0242	keep	.1498	allow	.2093
safe	system	operate	.2431	move	−.2363	start	.0013
right	accent	speak in	.1732	know	−.1223	hear	.0946
right	book	read	.1938	lend	−.0188	suggest	.0946
right	school	apply to	.2189	complain to	−.3736	reach	−.2756
right	structure	build	−.1084	teach	−.1084	support	−.0505
right	uniform	wear	.1990	provide	−.1084	look at	−.0505
wrong	author	accuse	−.1925	read	.0450	consider	.0653
wrong	colour	use	.2366	look for	−.0587	look at	−.1907
wrong	note	give	.0222	keep	−.2014	accept	−.1462
wrong	post	assume	−.3000	make	−.2579	consider	−.0466
wrong	strategy	adopt	.2804	encourage	.1937	look for	.0135

Table C.1: Materials for Experiment 8, with Mean Ratings (object interpretations)

Adjective	Noun	High		Medium		Low	
difficult	customer	buy	−.2682	pick	−.2050	begin	−.3560
difficult	friend	explain	−.4658	neglect	−.5274	enjoy	−.4711
difficult	passage	read	.1668	speak	−.3600	appear	−.4030
difficult	piece	read	.1052	survive	−.5080	continue	−.1006
difficult	spell	break	−.3047	create	−.2412	start	−.3661
easy	car	start	−.1652	move	−.2401	close	−.5750
easy	change	occur	.1999	prove	.2332	sit	−.0932
easy	food	cook	.0443	change	−.6046	form	−.3918
easy	habit	develop	.1099	start	.1156	appear	−.3490
easy	task	fit	−.3882	end	−.0982	continue	.0474
fast	device	go	.2638	come	−.3652	add	−.2219
fast	horse	run	.4594	work	.0025	add	−.5901
fast	lady	walk	−.0261	work	−.0716	see	−.4816
fast	pig	run	.2081	come	−.1807	get	−.2764
fast	town	grow	−.3601	spread	−.3289	sell	−.3462
good	ad	read	.1248	sell	.2154	run	−.0832
good	climate	change	−.3748	improve	−.3312	begin	−.4093
good	egg	look	−.0581	develop	−.2457	appear	.1149
good	light	work	−.0022	spread	−.2023	increase	−.4349
good	show	run	.0787	continue	−.0798	die	−.6569
hard	fish	bite	−.3583	pull	−.2579	appear	−.2568
hard	logic	get	−.1211	sell	−.4533	go	−.4388
hard	substance	keep	.0227	remain	.0978	seem	.1971
hard	toilet	flush	−.1796	look	−.3465	start	−.6835
hard	war	break out	−.4969	grow	−.2792	increase	−.2602
safe	building	approach	−.6815	stay	.0852	start	−.5152
safe	drug	come	−.3802	play	−.5562	try	−.3126
safe	man	eat	−.5434	ignore	−.6673	agree	−.6509
safe	speed	go	−.3116	leave	−.6136	remain	.1267
safe	system	operate	.3697	continue	.0374	think	−.4845
slow	child	react	.1485	adapt	.1556	express	−.0256
slow	hand	move	.0738	draw	.0039	work	−.0346
slow	meal	go	.1237	run	−.1474	become	−.2802
slow	minute	pass	.2717	start	−.4423	win	−.6709
slow	sleep	come	−.0671	follow	−.3108	seem	−.2169
right	accent	go	−.1727	sound	.1928	fall	−.3926
right	book	read	−.2429	feel	−.1027	discuss	−.2195
right	character	live	−.2505	set	−.4063	feel	−.1651
right	people	vote	−.4541	eat	−.5921	answer	−.0992
right	school	teach	.0159	start	−.3466	stand	−.3839
wrong	author	go	−.4348	think	−.4128	read	−.5542
wrong	business	think	−.4018	spend	−.4416	hope	−.5608
wrong	colour	go	.1846	show	−.0819	seem	.2869
wrong	note	conclude	−.2575	show	−.3480	tell	−.2732
wrong	policy	encourage	−.0401	identify	−.2167	accept	.0183

Table C.2: Materials for Experiment 8, with Mean Ratings (subject interpretations)

Bibliography

AAAI. 1996. *Proceedings of 13th National Conference on Artificial Intelligence*, Portland, OR.

Abney, Steve. 1996. Partial Parsing via Finite-State Cascades. In John Carroll, ed., *Workshop on Robust Parsing*, 8–15. European Summer School in Logic, Language and Information, Prague.

Abney, Steve. 1997. *The SCOL Manual*. University of Tübingen.

Abney, Steve, and Marc Light. 1999. Hiding a Semantic Class Hierarchy in a Markov Model. In Andrew Kehler and Andreas Stolcke, eds., *Proceedings of ACL Workshop on Unsupervised Learning in Natural Language Processing*, 1–8. College Park, MD.

ACL. 1984. *Proceedings of the 10th International Conference on Computational Linguistics*, Stanford, CA.

ACL. 1993. *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics*, Columbus, OH.

ACL. 1995. *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics*, Cambridge, MA.

ACL. 1996. *Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics*, Santa Cruz, CA.

ACL. 1999. *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, College Park, MD.

ACL/EACL. 1997. *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and 8th Conference of the European Chapter of the Association for Computational Linguistics*, Madrid, Spain.

Agirre, Eneko, and German Rigau. 1996. Word Sense Disambiguation Using Conceptual Density. In COLING (1996), 16–22.

Aho, Alfred V., Ravi Sethi, and Jeffrey D. Ullman. 1986. *Compilers: Principles, Techniques, and Tools*. Reading, MA: Addison-Wesley.

Anick, Peter, and James Pustejovsky. 1990. An Application of Lexical Semantics to Knowledge Acquisition from Corpora. In *Proceedings of 13th International Conference on Computational Linguistics*, 7–12. Helsinki, Finland.

Aone, Chinatsu, and Douglas McKee. 1995. Acquiring Predicate-Argument Mapping Information from Multilingual Texts. In Boguraev and Pustejovsky (1995a), 191–202.

Apresjan, Ju. D. 1973. Regular Polysemy. *Linguistics* 142: 5–32.

Aronoff, Mark. 1976. *Word Formation in Generative Grammar*. Linguistic Inquiry Monograph 1. Cambridge, MA: The MIT Press.

Asher, Nicholas, and Alex Lascarides. 1995. Lexical Disambiguation in a Discourse Context. *Journal of Semantics* 12(1): 69–108.

Atkins, Beryl T., and Beth Levin. 1991. Admitting Impediments. In Zernik (1991), 233–262.

Bard, Ellen Gurman, Dan Robertson, and Antonella Sorace. 1996. Magnitude Estimation of Linguistic Acceptability. *Language* 72(1): 32–68.

Bauer, Laurie. 1983. *English Word-formation*. Cambridge: Cambridge University Press.

Bergler, Sabine. 1991. The Semantics of Collocational Patterns for Reporting Verbs. In *Proceedings of the 5th Conference of the European Chapter of the Association for Computational Linguistics*, 216–221. Berlin, Germany.

Boguraev, Branimir K. 1979. Automatic Resolution of Word Sense Ambiguity. Ph.D. thesis, University of Cambridge.

Boguraev, Branimir K., Edward J. Briscoe, John Carroll, David M. Carter, and Claire Grover. 1987. The Derivation of a Grammatically-indexed Lexicon from the Longman Dictionary of Contemporary English. In *Proceedings of the 25th Annual Meeting of the Association for Computational Linguistics*, 193–200. Stanford, CA.

Boguraev, Branimir K., and Ted Briscoe. 1989. Utilising the LDOCE Grammar Codes. In Ted Briscoe and Branimir K. Boguraev, eds., *Computational Lexicography for Natural Language Processing*, 85–116. London: Longman.

Boguraev, Branimir K., and James Pustejovsky, eds. 1995a. *Corpus Processing for Lexical Acquisition*. Cambridge MA: The MIT Press.

Boguraev, Branimir K., and James Pustejovsky. 1995b. Issues in Text-Based Lexicon Acquisition. In Boguraev and Pustejovsky (1995a), 3–17.

Bouillon, Pierette. 1997. Polymorphie et Sémantique Lexicale: Le Cas des Adjectifs. Ph.D. thesis, Université de Paris 7 Denis Diderot.

Bourigault, Didier. 1992. Surface Grammatical Analysis for the Extraction of Terminological Noun Phrases. In COLING (1992), 977–981.

Bourigault, Didier, and Christian Jacquemin. 1999. Term Extraction and Term Clustering: An Integrated Platform for Computer Aided Terminology. In EACL (1999), 15–21.

Brent, Michael. 1993. From Grammar to Lexicon: Unsupervised Learning of Lexical Syntax. *Computational Linguistics* 19(3): 243–262.

Bresnan, Joan. 2000. *Lexical-Functional Syntax.* Oxford: Blackwell.

Brill, Eric. 1993. Transformation-Based Learning. Ph.D. thesis, University of Pennsylvania.

Brill, Eric, and Philip Resnik. 1994. A Rule-based Approach to Prepositional Phrase Attachment Disambiguation. In COLING (1994), 1198–1204.

Briscoe, Ted, and John Carroll. 1997. Automatic Extraction of Subcategorization from Corpora. In *Proceedings of the 5th Conference on Applied Natural Language Processing*, 356–363. Washington, DC.

Briscoe, Ted, and Ann Copestake. 1995. Dative Constructions as Lexical Rules in the TDFS Framework. ACQUILEX II Working Paper, Computer Laboratory, University of Cambridge.

Briscoe, Ted, and Ann Copestake. 1996. Controlling the Application of Lexical Rules. In Evelyn Viegas, ed., *Proceedings of ACL SIGLEX Workshop on Breadth and Depth of Semantic Lexicons*, 7–19. Santa Cruz, CA.

Briscoe, Ted, and Ann Copestake. 1999. Lexical Rules in Constraint-based Grammar. *Computational Linguistics* 25(4): 487–526.

Britton, B. K. 1978. Lexical Ambiguity of Words Used in English Text. *Behavior Research Methods and Instrumentation* 10: 1–7.

Bruce, Rebecca, and Janyce Wiebe. 1994. Word-sense Disambiguation Using Decomposable Models. In *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics*, 139–145. Las Cruces, NM.

Bruce, Rebecca, and Janyce Wiebe. 1999. Decomposable Modeling in Natural Language Processing. *Computational Linguistics* 25(2): 195–209.

Bryan, Robert M. 1973. Abstract Thesauri and Graph Theory Applications to Thesaurus Research. In Sally Yeates, ed., *Automated Language Analysis*, 45–89. Lawrence, KS: University of Kansas Press.

Buitelaar, Paul. 1997. A Lexicon for Underspecified Semantic Tagging. In SIGLEX (1997), 25–33.

Burnage, Gavin. 1990. *CELEX – A Guide for Users.* Centre for Lexical Information, University of Nijmegen.

Burnard, Lou. 1995. *Users Guide for the British National Corpus.* British National Corpus

Consortium, Oxford University Computing Service.

Carletta, Jean. 1996. Agreement on Classification Tasks: The Kappa Statistic. *Computational Linguistics* 22(2): 249–254.

Carroll, Glenn, and Mats Rooth. 1998. Valence Induction with a Head-lexicalized PCFG. In Nancy Ide and Atro Voutilainen, eds., *Proceedings of the 3rd Conference on Empirical Methods in Natural Language Processing*, 36–45. Granada, Spain.

Carroll, John, and Stephan Oepen. 2000. Efficient Large-Scale Parsing – A Survey. In *Proceedings of COLING Workshop on Efficiency in Large-scale Parsing Systems*, 7–11.

Chao, Gerald, and Michael G. Dyer. 2000. Word Sense Disambiguation of Adjectives Using Probabilistic Networks. In COLING (2000), 152–158.

Chierchia, Gennaro, and Sally McConnell-Ginet. 1990. *Meaning and Grammar: an Introduction to Semantics*. Cambridge, MA: The MIT Press.

Chomsky, Noam, and Morris Halle. 1968. *The Sound Pattern of English*. New York: Harper and Row.

Christ, Oliver. 1995. *The XKWIC User Manual*. Institute for Computational Linguistics, University of Stuttgart.

Church, Kenneth W. 1988. A Stochastic Parts Program and Noun Phrase Parser for Unrestricted Text. In *Proceedings of the 2nd Conference on Applied Natural Language Processing*, 136–143. Morristown, NJ.

Church, Kenneth W., and William A. Gale. 1991. A Comparison of the Enhanced Good-Turing and Deleted Estimation Methods for Estimating Probabilities of English Bigrams. *Computer Speech and Language* 5(1): 19–54.

Church, Kenneth W., and Patrick Hanks. 1990. Word Association Norms, Mutual Information, and Lexicography. *Computational Linguistics* 16(1): 22–29.

Church, Kenneth W., and Robert L. Mercer. 1993. Introduction to the Special Issue on Computational Linguistics Using Large Corpora. *Computational Linguistics* 19(1): 1–24.

Ciaramita, Massimiliano, and Mark Johnson. 2000. Explaining Away Ambiguity: Learning Verb Selectional Restrictions with Bayesian Networks. In COLING (2000), 187–193.

Clark, E. V., and H. H. Clark. 1979. When Nouns Surface as Verbs. *Language* 55(4): 767–811.

Cohen, J. 1960. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement* 20: 37–46.

Cohen, William W. 1996. Learning Trees and Rules with Set-valued Features. In AAAI (1996), 709–716.

COLING. 1992. *Proceedings of the 14th International Conference on Computational Linguistics*, Nantes, France.

COLING. 1994. *Proceedings of the 15th International Conference on Computational Linguistics*, Kyoto, Japan.

COLING. 1996. *Proceedings of the 16th International Conference on Computational Linguistics*, Copenhagen, Denmark.

COLING. 2000. *Proceedings of the 18th International Conference on Computational Linguistics*, Saarbrücken, Germany.

COLING/ACL. 1998. *Proceedings of the 17th International Conference on Computational Linguistics and 36th Annual Meeting of the Association for Computational Linguistics*, Montréal, Canada.

Collins, Michael. 1998. Head-driven Statistical Models for Natural Language Parsing. Ph.D. thesis, University of Pennsylvania.

Collins, Michael, and James Brooks. 1995. Prepositional Phrase Attachment through a Backed-Off Model. In David Yarowsky and Kenneth W. Church, eds., *Proceedings of the 3rd Workshop on Very Large Corpora*, 27–38. Cambridge, MA.

Copestake, Ann. 1992. The Representation of Lexical Semantic Information. Ph.D. thesis, University of Sussex.

Copestake, Ann. 1995. Representing Lexical Polysemy. In Judith Klavans, ed., *Proceedings of the AAAI Spring Symposium on Representation and Acquisition of Lexical Knowledge: Polysemy, Ambiguity and Generativity*, 21–26. Stanford, CA.

Copestake, Ann, and Ted Briscoe. 1992. Lexical Operations in a Unification-based Framework. In Pustejovsky and Bergler (1992), 88–95.

Copestake, Ann, and Alex Lascarides. 1997. Integrating Symbolic and Statistical Representations: The Lexicon Pragmatics Interface. In ACL/EACL (1997), 136–143.

Corley, Martin, and Margaret Cuthbert. 1997. Indiviual Differences in Modifier Attachments: Experience-Based Factors. Paper presented at the 3rd Conference on Architectures and Mechanisms for Language Processing, Edinburgh.

Corley, Martin, and Sarah Haywood. 1999. Parsing Modifiers: The Case of Bare-NP Adverbs. In *Proceedings of the 21st Annual Conference of the Cognitive Science Society*, 126–131. Mahwah, NJ: Lawrence Erlbaum Associates.

Corley, Martin, Frank Keller, and Christoph Scheepers. 2000. Conducting Psycholinguistic Experiments over the World Wide Web. Unpubl. ms., University of Edinburgh and Saarland University.

Corley, Steffan, Martin Corley, Frank Keller, Matthew W. Crocker, and Shari Trewin. 2001. Finding Syntactic Structure in Unparsed Corpora: The Gsearch Corpus Query System. *Computers and the Humanities* 35(2): 81–94.

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

Cowie, A P., ed. 1989. *Oxford Advanced Learner's Dictionary*. Oxford: Oxford University Press.

Dagan, Ido, Lilian Lee, and Fernando C. N. Pereira. 1999. Similarity-Based Models of Word Cooccurrence Probabilities. *Machine Learning* 34(1–3): 43–69.

Daille, Béatrice. 1996. Study and Implementation of Combined Techniques for Automatic Extraction of Terminology. In Judith Klavans and Philip Resnik, eds., *The Balancing Act: Combining Symbolic and Statistical Approaches to Language*, 49–66. Cambridge, MA: The MIT Press.

Dang, Hoa Trang, Karin Kipper, Martha Palmer, and Joseph Rosenzweig. 1998. Investigating Regular Sense Extensions Based on Intersective Levin Classes. In COLING/ACL (1998), 293–299.

Dang, Hoa Trang, Joseph Rosenzweig, and Martha Palmer. 1997. Associating Semantic Components with Intersective Levin Classes. In *Proceedings of the 1st AMTA SIG-IL Workshop on Interlinguas*, 1–8. San Diego, CA.

Dempster, A. P., N. M. Laird, and D. B. Rubin. 1977. Maximum Likelihood for Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society* 39(2): 1–38.

Dini, Luca, Vittorio Di Tomaso, and Frédérique Segond. 1998. Error-driven Word Sense Disambiguation. In COLING/ACL (1998), 320–324.

Dixon, R.M.W. 1991. *A New Approach to English Grammar on Semantic Principles*. Oxford: Oxford University Press.

Dorr, Bonnie J. 1997. Large-Scale Dictionary Construction for Foreign Language Tutoring and Interlingual Machine Translation. *Machine Translation* 12(4): 371–322.

Dorr, Bonnie J., and Doug Jones. 1996. Role of Word Sense Disambiguation in Lexical Acquisition: Predicting Semantics from Syntactic Cues. In COLING (1996), 322–327.

Downing, Pamela. 1977. On the Creation and Use of English Compound Nouns. *Language* 53(4): 810–842.

Dunning, Ted. 1993. Accurate Methods for the Statistics of Surprise and Coincidence. *Computational Linguistics* 19(1): 61–74.

EACL. 1999. *Proceedings of the 9th Conference of the European Chapter of the Association*

for Computational Linguistics, Bergen, Norway.

Earley, Jay. 1970. An Efficient Context-Free Parsing Algorithm. *Communications of the ACM* 13(2): 94–102.

Edward, Kelly, and Philip Stone. 1975. *Computer Recognition of English Word Senses*. Amsterdam: North Holland.

Elworthy, David. 1994. Does Baum-Welch Re-estimation Help Taggers? In *Proceedings of the 4th Conference on Applied Natural Language Processing*, 53–58. Stuttgart, Germany.

EMNLP. 1999. *Proceedings of Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, College Park, MD.

Essen, Ute, and Volker Steinbiss. 1992. Co-occurrence Smoothing for Stochastic Language Modeling. In *Proceedings of International Conference on Acoustics Speech and Signal Processing*, 161–164. San Francisco, CA.

Fano, Robert. 1961. *Transmission of Information: A Statistical Theory of Communications*. Cambridge, MA: The MIT Press.

Finch, Steve. 1993. Finding Structure in Language. Ph.D. thesis, University of Edinburgh.

Finin, Tim. 1980. The Semantic Interpretation of Nominal Compounds. In *Proceedings of 1st National Conference on Artificial Intelligence*, 310–315. Stanford, CA.

Gale, William A., Kenneth W. Church, and David Yarowsky. 1992. A Method for Disambiguating Word Senses in a Large Corpus. *Computers and the Humanities* 26(5–6): 415–439.

Goldberg, Adele. 1995. *Constructions*. Chicago: Chicago University Press.

Grefenstette, Gregory. 1994. *Explorations in Automatic Thesaurus Discovery*. Boston London: Kluwer Academic Publishers.

Grefenstette, Gregory, and Simone Teufel. 1995. Corpus-Based Method for Automatic Identification of Support Verbs for Nominalizations. In *Proceedings of the 7th Conference of the European Chapter of the Association for Computational Linguistics*, 98–103. Dublin.

Grishman, Ralph, Catherine Macleod, and Adam Meyers. 1994. COMLEX Syntax: Building a Computational Lexicon. In COLING (1994), 268–272.

Grishman, Ralph, and John Sterling. 1994. Generalizing Automatically Generated Selectional Patterns. In COLING (1994), 742–747.

Guthrie, Joe A., Louise Guthrie, Yorick Wilks, and Brian M. Slator. 1991. Subject-dependent Co-occurrence and Word Sense Disambiguation. In *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics*, 146–152. Berkeley, CA.

Hatzivassiloglou, Vasileios, and Kathy McKeown. 1995a. A Quantitative Evaluation of Lin-

guistic Tests for the Automatic Prediction of Semantic Markedness. In ACL (1995), 197–204.

Hatzivassiloglou, Vasileios, and Kathy McKeown. 1995b. Towards the Automatic Identification of Adjectival Scales: Clustering Adjectives According to Meaning. In ACL (1993), 172–182.

Hatzivassiloglou, Vasileios, and Kathy McKeown. 1997. Predicting the Semantic Orientation of Adjectives. In ACL/EACL (1997), 174–181.

Hatzivassiloglou, Vasileios, and Janyce Wiebe. 2000. Effects of Adjective Orientation and Gradability on Sentence Subjectivity. In COLING (2000), 299–305.

Hatzivassiloglou, Vassilis, Judith L. Klavans, and Eleazar Eskin. 1999. An Annotation Scheme for Discourse-level Argumentation in Research Articles. In EMNLP (1999), 203–212.

Hearst, Marti. 1991. Noun Homograph Disambiguation. In *Proceedings of the 7th Annual Conference of the University of Waterloo Centre for the New OED and Text Research*, 1–19. Ontario, Canada.

Hearst, Marti. 1992. Automatic Acquisition of Hyponyms from Large Text Corpora. In COLING (1992), 539–545.

Hindle, Donald. 1989. Acquiring Disambiguation Rules from Text. In *Proceedings of 27th Annual Meeting of the Association for Computational Linguistics*, 118–125. Vancouver, Canada.

Hindle, Donald. 1990. Noun Classification from Predicate-Argument Structures. In *Proceedings of the 28th Annual Meeting of the Association for Computational Linguistics*, 268–275. Pittsburgh, PA.

Hindle, Donald, and Mats Rooth. 1993. Structural Ambiguity and Lexical Relations. *Computational Linguistics* 19(1): 103–120.

Hirschberg, Julia, and Christine Nakatani. 1996. A Prosodic Analysis of Discourse Segments in Direction-giving Monologues. In ACL (1996), 286–293.

Hirst, Graeme. 1987. *Semantic Interpretation and the Resolution of Ambiguity*. Cambridge: Cambridge University Press.

Hobbs, Jerry R., Mark Stickel, Douglas Appelt, and Paul Martin. 1993. Interpretation as Abduction. *Journal of Artificial Intelligence* 63(1–2): 69–142.

Hornby, Albert Sydney. 1989. *Oxford Advanced Learner's Dictionary of Current English*. Oxford: Oxford University Press.

Howes, D. H., and R. L. Solomon. 1951. Visual Duration Threshold as a Function of Word Probability. *Journal of Experimental Psychology* 41: 401–410.

Ide, Nancy, and Jean Véronis. 1998. Introduction to the Special Issue on Word Sense Disambiguation: The State of the Art. *Computational Linguistics* 24(1): 1–40.

Isabelle, Pierre. 1984. Another Look at Nominal Compounds. In ACL (1984), 509–516.

Jackendoff, Ray. 1975. Morphological and Semantic Regularities in the Lexicon. *Language* 51(3): 639–71.

Jackendoff, Ray. 1990. *Semantic Structures*. Cambridge, MA: The MIT Press.

Jacquemin, Christian. 1996. A Symbolic and Surgical Acquisition of Terms Through Variation. In Stefan Wermter, Ellen Riloff, and Gabriele Scheler, eds., *Connectionist, Statistical and Symbolic Approaches to Learning for Natural Language*, Lecture Notes in Artificial Intelligence, 425–438. Springer, Berlin.

Jones, Bernard. 1995. Predicating Nominal Compounds. In *Proceedings of 17th Annual Conference of the Cognitive Science Society*, 130–135. Pittsburgh, PA.

Jurafsky, Daniel. 1996. A Probabilistic Model of Lexical and Syntactic Access and Disambiguation. *Cognitive Science* 20: 137–194.

Jurafsky, Daniel, and James H. Martin. 2000. *Speech and Language Processing*. Upper Saddle River, NJ: Prentice Hall.

Justeson, John S., and Slava M. Katz. 1995a. Co-occurrences of Antonymous Adjectives and their Contexts. *Computational Linguistics* 17(1): 1–19.

Justeson, John S., and Slava M. Katz. 1995b. Principled Disambiguation: Discriminating Adjective Senses with Modified Nouns. *Computational Linguistics* 21(1): 1–27.

Justeson, John S., and Slava M. Katz. 1995c. Technical Terminology: Some Linguistic Properties and an Algorithm for Identification in Text. *Natural Language Engineering* 1(1): 9–27.

Karp, Daniel, Yves Schabes, Martin Zaidel, and Dania Egedi. 1992. A Freely Available Wide Coverage Morphological Analyzer for English. In COLING (1992), 950–954.

Katz, Slava M. 1987. Estimation of Probabilities from Sparse Data for the Language Model Component of a Speech Recognizer. *IEEE Transactions on Acoustics Speech and Signal Processing* 33(3): 400–401.

Keller, Frank. 2000. Gradience in Grammar: Experimental and Computational Aspects of Degrees of Grammaticality. Ph.D. thesis, University of Edinburgh.

Keller, Frank, and Theodora Alexopoulou. 2001. Phonology Competes with Syntax: Experimental Evidence for the Interaction of Word Order and Accent Placement in the Realization of Information Structure. *Cognition* To appear.

Keller, Frank, Martin Corley, Steffan Corley, Lars Konieczny, and Amalia Todirascu. 1998.

WebExp: A Java Toolbox for Web-Based Psychological Experiments. Technical Report HCRC/TR-99, Human Communication Research Centre, University of Edinburgh.

Kilgariff, Adam. 1992. Polysemy. Ph.D. thesis, University of Sussex.

Korhonen, Anna. 1998. Automatic Extraction of Subcategorization Frames for Corpora – Improving Filtering with Diathesis Alternations. In *Workshop on Automated Acquisition of Syntax and Parsing*, 49–56. European Summer School in Logic, Language and Information, Saarbrücken, Germany.

Kupiec, Julian. 1992. Robust Part-of-speech Tagging Using a Hidden Markov Model. *Computer Speech and Language* 6(3): 225–242.

Lahav, Ran. 1989. Against Compositionality: The Case of Adjectives. *Philosophical Studies* 57: 261–279.

Lakoff, George, and Mark Johnson. 1980. *Metaphors We Live By*. Chicago: University of Chicago Press.

Landauer, Thomas K., and Susan T. Dumais. 1997. A Solution to Plato's Problem: The Latent Semantic Analysis Theory of Acquisition, Induction and Representation of Knowledge. *Psychological Review* 104(2): 211–240.

Landis, J. R., and G. G. Koch. 1977. The Measurement of Observer Agreement for Categorical Data. *Biometrics* 33: 159–174.

Lapata, Maria. 1999a. Acquiring Lexical Generalizations from Corpora: A Case Study for Diathesis Alternations. In ACL (1999), 397–404.

Lapata, Maria. 1999b. The Acquisition and Interpretation of Compound Nouns. *Computational Linguistics* Submitted.

Lapata, Maria. 1999c. Corpus-Based Induction of Lexical Representation and Meaning. In *Proceedings of the 16th National Conference on Artificial Intelligence*, 948. Orlando, FL.

Lapata, Maria. 2000. The Automatic Interpretation of Nominalizations. In *Proceedings of the 17th National Conference on Artificial Intelligence*, 716–721. Austin, TX.

Lapata, Maria. 2001. A Corpus-based Account of Regular Polysemy: The Case of Context-sensitive Adjectives. In *Proceedings of the 2nd North American Annual Meeting of the Association for Computational Linguistics*. Pittsburgh, PA.

Lapata, Maria, and Chris Brew. 1999. Using Subcategorization to Resolve Verb Class Ambiguity. In EMNLP (1999), 266–274.

Lapata, Maria, Frank Keller, and Sabine Schulte im Walde. 2001. Verb Frame Frequency as a Predictor of Verb Bias. *Journal of Psycholinguistic Research* To appear.

Lapata, Maria, Scott McDonald, and Frank Keller. 1999. Determinants of Adjective-Noun

Plausibility. In EACL (1999), 30–36.

Lascarides, Alex. 1995. The Pragmatics of Word Meaning. In *Proceedings of the AAAI Spring Symposium Series: Representation and Acquisition of Lexical Knowledge: Pulysemy, Ambiguity and Generativity*, 75–80. Stanford, CA.

Lascarides, Alex, and Ann Copestake. 1998. Pragmatics and Word Meaning. *Journal of Linguistics* 34(2): 387–414.

Lauer, Mark. 1995. Designing Statistical Language Learners: Experiments on Compound Nouns. Ph.D. thesis, Macquarie University.

Leacock, Claudia, Martin Chodorow, and George A. Miller. 1998. Using Corpus Statistics and WordNet Relations for Sense Identification. *Computational Linguistics* 14(1): 147–165.

Lee, Lilian. 1999. Measures of Distributional Similarity. In ACL (1999), 25–32.

Leech, Geoffrey, Roger Garside, and Michael Bryant. 1994. The Tagging of the British National Corpus. In COLING (1994), 622–628.

Leonard, Rosemary. 1984. *The Interpretation of English Noun Sequences on the Computer*. Amsterdam: North-Holland.

Lesk, Michael. 1986. Automatic Sense Disambiguation: How to Tell a Pine Cone from an Ice Cream Cone. In *Proceedings of the 1986 Special Interest Group in Documentation*, 24–26. New York: Association for Computing Machinery.

Levi, Judith N. 1978. *The Syntax and Semantics of Complex Nominals*. New York: Academic Press.

Levin, Beth. 1993. *English Verb Classes and Alternations: A Preliminary Investigation*. Chicago: University of Chicago Press.

Li, Hang, and Naoki Abe. 1995. Generalizing Case Frames Using a Thesaurus and the MDL Principle. In *Proceedings of 1st International Conference on Recent Advances in Natural Language Processing*, 239–248. Tzigov Chark, Bulgaria.

Liberman, Mark, and Richard Sproat. 1992. The Stress and Structure of Modified Noun Phrases in English. In Ivan Sag and Ann Szabolcsi, eds., *Lexical Matters*, 131–181. Stanford, CA: CSLI Publications.

Light, Mark. 1996. Morphological Cues for Lexical Semantics. In ACL (1996), 25–31.

Lin, Dekang. 1999. Automatic Identification of Non-Compositional Phrases. In ACL (1999), 317–324.

Lodge, Milton. 1981. *Magnitude Scaling: Quantitative Measurement of Opinions*. Beverley Hills, CA: Sage Publications.

Lyons, John. 1977. *Semantics*. Cambridge: Cambridge University Press.

Macleod, Catherine, Ralph Grishman, Adam Meyers, Leslie Barrett, and Ruth Reeves. 1998. NOMLEX: A Lexicon of Nominalizations. In *Proceedings of the 8th International Congress of the European Association for Lexicography*, 187–193. Liège, Belgium.

Manning, Christopher D. 1993. Automatic Acquisition of a Large Subcategorization Dictionary from Corpora. In ACL (1993), 235–242.

Manning, Christopher D., and Hinrich Schütze. 1999. *Foundations of Statistical Natural Language Processing*. Cambridge, MA: The MIT Press.

Marcus, Mitchell, Kim Grace, Mary Ann Marcinkiewicz, Robert MacIntyre, Ann Bies, Mark Ferguson, Karen Katz, and Britta Schasberger. 1994. The Penn Treebank: Annotating Predicate Argument Structure. In *Proceedings of Human Language Technology Workshop*, 110–115. Morgan Kaufman.

Marcus, Mitchell P., Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a Large Annotated Corpus of English: The Penn Treebank. *Computational Linguistics* 19(2): 313–330.

Marsh, Elaine. 1984. A Computational Analysis of Complex Noun Phrases in Navy Messages. In ACL (1984), 505–508.

Masterman, Margaret. 1957. The Thesaurus in Syntax and Semantics. *Mechanical Translation* 4: 1–2.

McCarthy, Diana. 2000. Using Semantic Preferences to Identify Verbal Participation in Role Switching Alternations. In *Proceedings of the 1st North American Annual Meeting of the Association for Computational Linguistics*, 256–263. Seattle, WA.

McCarthy, Diana, and Anna Korhonen. 1998. Detecting Verbal Participation in Diathesis Alternations. In COLING/ACL (1998), 1493–1495. Student Session.

McDonald, David. 1982. Understanding Noun Compounds. Ph.D. thesis, Carnegie Mellon University.

McDonald, Scott. 2000. Environmental Determinants of Lexical Processing Effort. Ph.D. thesis, University of Edinburgh.

Merlo, Paola, and Suzanne Stevenson. 1999. Automatic Verb Classification Using Distributions of Grammatical Features. In EACL (1999), 45–51.

Mihalcea, Rada, and Dan Moldovan. 1998. Word Sense Disambiguation Based on Semantic Density. In Sanda Harabagiu, ed., *Proceedings of COLING/ACL Workshop on Usage of WordNet in Natural Language Processing*, 16–22. Montréal, Canada.

Miller, George A., Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine J.

Miller. 1990. Introduction to WordNet: An On-line Lexical Database. *International Journal of Lexicography* 3(4): 235–244.

Miller, George A., and William G. Charles. 1991. Contextual Correlates of Semantic Similarity. *Language and Cognitive Processes* 6(1): 1–28.

Monsell, S. 1991. The Nature and Locus of Word Frequency Effects in Reading. In D. Besner and G. W. Humphreys, eds., *Basic Processes in Reading: Visual Word Recognition*. Hillsdale, NJ: Erlbaum.

Mosteller, Frederick, and David L. Wallace. 1964. *Inference and Disputed Authorship: The Federalist*. London: Addison-Wesley.

Narayanan, Srini, and Daniel Jurafsky. 1998. Bayesian Models of Human Sentence Processing. In *Proceedings of the 20th Annual Conference of the Cognitive Science Society*. Mahwah, NJ: Lawrence Erlbaum Associates.

Ng, Tou Hwee, and Hian Beng Lee. 1996. Integrating Multiple Knowledge Sources to Disambiguate Word Senses: An Exemplar-based Approach. In ACL (1996), 40–47.

Ostler, Nicholas, and B. T. S. Atkins. 1992. Predictable Meaning Shift: Some Linguistic Properties of Lexical Implication Rules. In Pustejovsky and Bergler (1992), 87–100.

Palmer, Martha. 2000. Consistent Criteria for Sense Distinctions. *Computers and the Humanities* 34(1–2): 217–222.

Passonneau, Rebecca J., and Diane J. Litman. 1997. Discourse Segmentation by Human and Automated Means. *Computational Linguistics* 23(1): 103–140.

Patel, Malti, John A. Bullinaria, and Joseph P. Levy. 1998. Extracting Semantic Representations from Large Text Corpora. In John A. Bullinaria, D. W. Glasspool, and G. Houghton, eds., *In Proceedings of the 4th Workshop on Neural Computation and Psychology*, 199–212. Springer, Berlin.

Pedersen, Ted, Mehmet Kayaalp, and Rebecca Bruce. 1996. Significant Lexical Relationships. In AAAI (1996), 455–460.

Pinker, Steven. 1989. *Learnability and Cognition: The Acquisition of Argument Structure*. Cambridge, MA: The MIT Press.

Pollard, Carl, and Ivan A. Sag. 1987. *Information-Based Syntax and Semantics*. No. 13 in CSLI Lecture Notes. Stanford, CA: CSLI Publications.

Pollard, Carl, and Ivan A. Sag. 1994. *Head-Driven Phrase Structure Grammar*. Chicago, IL and Stanford, CA: University of Chicago Press and CSLI Publications.

Poznański, Victor, and Antonio Sanfilippo. 1995. Detecting Dependencies between Semantic Verb Classes. In Boguraev and Pustejovsky (1995a), 175–190.

Procter, Paul. 1978. *Longman Dictionary of Contemporary English.* London: Longman.

Pustejovsky, James. 1995. *The Generative Lexicon.* Cambridge, MA: The MIT Press.

Pustejovsky, James, and Sabine Bergler, eds. 1992. *Lexical Semantics and Knowledge Representation.* New York: Springer.

Pustejovsky, James, Sabine Bergler, and Peter Anick. 1993. Lexical Semantic Techniques for Corpus Analysis. *Computational Linguistics* 19(3): 331–358.

Quirk, Randolph, Sidney Greenbaum, Geoffrey Leech, and Jan Svartvik. 1985. *A Comprehensive Grammar of the English Language.* London: Longman.

Ratnaparkhi, Adwait. 1998. Unsupervised Statistical Models for Prepositional Phrase Attachment. In COLING/ACL (1998), 1079–1085.

Redington, M., N. Chater, and S. Finch. 1998. Distributional Information: A Powerful Cue for Acquiring Syntactic Categories. *Cognitive Science* 22(4): 425–469.

Resnik, Philip. 1997. Selectional Preferences and Sense Disambiguation. In SIGLEX (1997), 52–57.

Resnik, Philip, and Mona Diab. 2000. Measuring Verb Similarity. In *Proceedings of the 22nd Annual Conference of the Cognitive Science Society*, 399–404. Mahwah, NJ: Lawrence Erlbaum Associates.

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

Roland, Douglas, and Daniel Jurafsky. 2000. Verb Sense and Verb Subcategorization Probabilities. In Paola Merlo and Suzanne Stevenson, eds., *The Lexical Basis of Human Sentence Processing: Formal and Computational Issues.* Amsterdam: John Bejamins. To appear.

Rooth, Mats, Stefan Riezler, Detlef Prescher, Glenn Carroll, and Franz Beil. 1999. Inducing a Semantically Annotated Lexicon via EM-Based Clustering. In ACL (1999), 104–111.

Russell, Stuart J., and Peter Norvig. 1995. *Artifical Intelligence: A Modern Approach.* New Jersey: Prentice Hall.

Saffran, J. R., R. N. Aslin, and E. L. Newport. 1996. Statistical Learning by 8-Month Old Infants. *Science* 274: 1926–1928.

Sapir, Edward. 1944. Grading: A Study in Semantics. *Philosophy of Science* 11: 83–116.

Schooler, L. J., and J. R. Anderson. 1997. The Role of Process in the Rational Analysis of Memory. *Cognitive Psychology* 32(3): 219–250.

Schulte im Walde, Sabine. 1998. Automatic Semantic Classification of Verbs According to their Alternation Behaviour. Master's thesis, Institut für Maschinelle Sprachverarbeitung,

University of Stuttgart.

Schulte im Walde, Sabine. 2000. Clustering Verbs Semantically According to their Alternation Behaviour. In COLING (2000), 747–753.

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

Schütze, Hinrich. 1993. Word Space. In Stephen José Hanson, Jack D. Cowan, and C. Lee Giles, eds., *Advances in Neural Information Processing Systems*, 895–902. San Mateo, CA: Morgan Kaufmann.

Schütze, Hinrich. 1998. Automatic Word Sense Discrimination. *Computational Linguistics* 24(1): 97–124.

Selkirk, Elizabeth. 1982. *The Syntax of Words*. Cambridge, MA: The MIT Press.

Shannon, Claude E., and Warren Weaver. 1949. *The Mathematical Theory of Communication*. Urbana: University of Illinois Press.

Siegel, Eric V. 1999. Corpus-Based Linguistic Indicators for Aspectual Classification. In ACL (1999), 112–119.

Siegel, Eric V., and Kathleen R. McKeown. 1994. Emergent Linguistic Rules from Inducing Decision Trees: Disambiguating Discourse Clue Words. In *Proceedings of the 12th National Conference on Artificial Intelligence*, 820–826. Seattle, WA.

Siegel, Sidney, and N. John Castellan. 1988. *Non Parametric Statistics for the Behavioral Sciences*. New York: McGraw-Hill.

SIGLEX. 1997. *Proceedings of the ACL SIGLEX Workshop on Tagging Text with Lexical Semantics: Why, What, and How?*, Washington, DC.

Smadja, F. 1992. XTRACT: An Overview. *Computers and the Humanities* 26(5–6): 399–414.

Smadja, Frank. 1991. Macrocoding the Lexicon with Co-occurrence Knowledge. In Zernik (1991), 165–189.

Small, Rieger. 1980. Word Expert Parsing: A Theory of Distributed Word-Based Natural Language Understanding. Ph.D. thesis, University of Maryland.

Sparck Jones, Karen. 1983. Compound Noun Interpretation Problems. Technical Report 45, Computer Laboratory, University of Cambridge.

Spence, D. P., and K. C. Owens. 1990. Lexical Co-occurence and Association Strength. *Journal of Psycholinguistic Research* 19(5): 317–330.

Spencer, Andrew. 1991. *Morphological Theory*. Oxford: Blackwell Publishers.

Spivey-Knowlton, Michael, and Julie C. Sedivy. 1995. Resolving Attachment Ambiguities

with Multiple Constraints. *Cognition* 55(3): 227–267.

Sproat, Richard. 1994. English Noun-Phrase Accent Prediction for Text-to-Speech. *Computer Speech and Language* 8(2): 79–94.

Stede, Manfred. 1998. A Generative Perspective on Verb Alternations. *Computational Linguistics* 24(3): 401–430.

Steedman, Mark. 2000. *The Syntactic Process*. Cambridge, MA: The MIT Press.

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

Stevenson, Suzanne, and Paola Merlo. 2000. Automatic Lexical Acquisition Based on Statistical Distributions. In COLING (2000), 815–821.

Strzalkowski, Tomek, and Barbara Vauthey. 1992. Information Retrieval Using Robust Natural Language Processing. In *Proceedings of the 30th Annual Meeting of the Association for Computational Linguistics*, 104–111. Columbus, OH.

Sturt, Patrick, Martin J. Pickering, and Matthew W. Crocker. 1999. Structural Change and Reanalysis Difficulty in Language Comprehension. *Journal of Memory and Language* 40(1): 136–150.

Sussna, Michael. 1993. Word Sense Disambiguation for Free-text Indexing Using a Massive Semantic Network. In *Proceedings of the 2nd International Conference on Information and Knowledge Management*, 67–74. Arlington, VA.

Teufel, Simone. 2000. Argumentative Zoning: Information Extraction from Scientific Articles. Ph.D. thesis, University of Edinburgh.

Vanderwende, Lucy. 1994. Algorithm for Automatic Interpretation of Noun Sequences. In COLING (1994), 782–788.

Vendler, Zeno. 1968. *Adjectives and Nominalizations*. The Hague: Mouton.

Voorhees, Ellen M. 1993. Using WordNet to Disambiguate Word Senses for Text Retrieval. In *Proceedings of the 16th ACM Conference on Research and Development in Information Retrieval*, 171–180. Pittsburgh, PA.

Warren, Beatrice. 1978. *Semantic Patterns of Noun-Noun Compounds*. Göteborg: Acta Universitatis Gothoburgensis.

Warren, Beatrice. 1984. *Classifying Adjectives*. Göteborg: Acta Universitatis Gothoburgensis.

Wechsler, Stephen. 1995. *The Semantic Basis of Argument Structure*. Stanford, CA: CSLI Publications.

Weeber, Marc, Rein Vos, and Harald R. Baayen. 2000. Extracting the Lowest-Frequency

Words: Pitfalls and Possibilities. *Computational Linguistics* 26(3): 301–317.

Weiss, Sholom M., and Casimir A. Kulikowski. 1991. *Computer Systems that Learn: Classification and Prediction Methods from Statistics, Neural Nets, Machine Learning, and Expert Systems*. San Mateo, CA: Morgan Kaufmann.

Whaley, C. P. 1978. Word-nonword Classification Time. *Journal of Verbal Learning and Verbal Behavior* 17: 143–154.

Whittaker, Joe. 1990. *Graphical Models in Applied Multivariate Statistics*. New York: John Wiley & Sons.

Wiebe, Janyce M., Thomas P. O'Hara, Thorsten Öhrström Sandgren, and Kenneth J. McKeever. 1998. An Empirical Approach to Temporal Reference Resolution. *Journal of Artifical Intelligence Research* 9: 247–293.

Wilks, Yorick. 1975. A Preferential, Pattern-seeking Semantics for Natural Language Inference. *Artificial Intelligence* 6(1): 53–74.

Wilks, Yorick, Brian M. Slator, and Louise Guthrie. 1996. *Electric Words: Dictionaries, Computers, and Meanings*. Cambridge, MA: The MIT Press.

Wilks, Yorick, and Mark Stevenson. 1998. Word Sense Disambiguation Using Optimized Combinations of Knowledge Sources. In COLING/ACL (1998), 1398–1092.

Wu, Dekai. 1993. Approximating Maximum-Entropy Ratings for Evidential Parsing and Semantic Interpretation. In *Proceedings of 13th International Joint Conference on Artificial Intelligence*, 1290–1296. Chamberry, France.

Yarowsky, David. 1992. Word Sense Disambiguation Using Statistical Models of Roget's Categories Trained on Large Corpora. In COLING (1992), 450–460.

Yarowsky, David. 1993. One Sense per Collocation. In *Proceedings of the ARPA Human Language Technology Workshop*, 266–271. Princeton, NJ.

Yarowsky, David. 1995. Unsupervised Word Sense Disambiguation Rivaling Supervised Methods. In ACL (1995), 189–196.

Zamparelli, Roberto. 1998. A Theory of Kinds, Partitives and of/z Possessives. In Chris Wilder and Artemis Alexiadou, eds., *Possessors, Predicates and Movement in the Determiner Phrase*, 259–301. Amsterdam: John Benjamins.

Zernik, Uri, ed. 1991. *Lexical Acquisition: Using Online Resources to Build a Lexicon*. Hillsdale, NJ: Erlbaum.

Index of Citations

Abney and Light (1999), 26, 38
Abney (1996), 36–38, 146, 166, 214
Abney (1997), 102
Agirre and Rigau (1996), 206
Aho et al. (1986), 37
Anick and Pustejovsky (1990), 21
Aone and McKee (1995), 15
Apresjan (1973), 17, 241
Aronoff (1976), 72, 74
Asher and Lascarides (1995), 21
Atkins and Levin (1991), 23
Bard et al. (1996), 46, 156, 158
Bauer (1983), 231
Bergler (1991), 21
Boguraev and Briscoe (1989), 15, 51, 74
Boguraev and Pustejovsky (1995b), 15
Boguraev et al. (1987), 105
Boguraev (1979), 22
Bouillon (1997), 141
Bourigault and Jacquemin (1999), 175, 176
Bourigault (1992), 175, 176
Brent (1993), 26, 68, 104–106
Bresnan (2000), 73
Brill and Resnik (1994), 61, 206
Brill (1993), 15
Briscoe and Carroll (1997), 26, 68, 69, 104–107, 122
Briscoe and Copestake (1995), 74
Briscoe and Copestake (1996), 20
Briscoe and Copestake (1999), 19–21, 51, 72–74, 243
Britton (1978), 22
Bruce and Wiebe (1994), 26
Bruce and Wiebe (1999), 43
Bryan (1973), 23
Buitelaar (1997), 24
Burnage (1990), 33, 34, 154, 182, 214
Burnard (1995), 31, 51
Carletta (1996), 45
Carroll and Oepen (2000), 241
Carroll and Rooth (1998), 105, 122, 241
Chao and Dyer (2000), 23, 171
Chierchia and McConnell-Ginet (1990), 139
Chomsky and Halle (1968), 174
Christ (1995), 179, 185, 187, 189
Church and Gale (1991), 216
Church and Hanks (1990), 57, 175–177, 181, 183, 205
Church and Mercer (1993), 177
Church (1988), 32
Ciaramita and Johnson (2000), 206
Clark and Clark (1979), 20
Cohen (1960), 45, 59, 127, 196, 221
Cohen (1996), 196, 227
Collins and Brooks (1995), 61, 216
Collins (1998), 38, 241
Copestake and Briscoe (1992), 19
Copestake and Lascarides (1997), 19, 25, 74, 193, 231
Copestake (1992), 19, 242
Copestake (1995), 20
Corley and Cuthbert (1997), 36
Corley and Haywood (1999), 36

Corley et al. (2000), 157
Corley et al. (2001), 36, 54, 63, 78, 118, 178
Cowart (1997), 46, 156
Cowie (1989), 23
Dagan et al. (1999), 38, 215, 218
Daille (1996), 15, 57, 175–177, 180, 181, 183, 189, 205
Dang et al. (1997), 104, 134
Dang et al. (1998), 113
Dempster et al. (1977), 106
Dini et al. (1998), 23
Dixon (1991), 77, 88
Dorr and Jones (1996), 133–136
Dorr (1997), 52, 104, 107, 134
Downing (1977), 174–176, 183, 207, 211, 235
Dunning (1993), 57, 175, 181, 189, 205
Earley (1970), 36
Edward and Stone (1975), 22
Elworthy (1994), 105
Essen and Steinbiss (1992), 218
Fano (1961), 181
Finch (1993), 15
Finin (1980), 210–212, 233
Gale et al. (1992), 22
Goldberg (1995), 51
Grefenstette and Teufel (1995), 15, 26
Grefenstette (1994), 15
Grishman and Sterling (1994), 215, 218, 219
Grishman et al. (1994), 35, 69, 84, 105
Guthrie et al. (1991), 23
Hatzivassiloglou and McKeown (1995a), 26, 170, 171, 206, 234
Hatzivassiloglou and McKeown (1995b), 170
Hatzivassiloglou and McKeown (1997), 170, 171, 206, 234
Hatzivassiloglou and Wiebe (2000), 170, 171, 206, 234
Hatzivassiloglou et al. (1999), 45
Hearst (1991), 23, 26, 132
Hearst (1992), 26, 206
Hindle and Rooth (1993), 55, 61–63, 65, 106, 118, 122, 213
Hindle (1989), 32
Hindle (1990), 15
Hirschberg and Nakatani (1996), 45
Hirst (1987), 22
Hobbs et al. (1993), 211, 234
Hornby (1989), 105
Howes and Solomon (1951), 244
Ide and Véronis (1998), 22–24
Isabelle (1984), 210
Jackendoff (1975), 74
Jackendoff (1990), 60
Jacquemin (1996), 175, 176, 205, 206
Jones (1995), 193, 233
Jurafsky and Martin (2000), 22
Jurafsky (1996), 244
Justeson and Katz (1995a), 15, 23
Justeson and Katz (1995b), 170, 171
Justeson and Katz (1995c), 175–177, 180, 189, 205
Karp et al. (1992), 32
Katz (1987), 215, 216
Keller and Alexopoulou (2001), 157
Keller et al. (1998), 157
Keller (2000), 46, 156
Kilgariff (1992), 20, 23, 24
Korhonen (1998), 104, 107
Kupiec (1992), 105, 120
Lahav (1989), 140, 141
Lakoff and Johnson (1980), 231

Landauer and Dumais (1997), 133
Landis and Koch (1977), 45
Lapata and Brew (1999), 29
Lapata et al. (1999), 36
Lapata et al. (2001), 36, 244
Lapata (1999a), 29
Lapata (1999b), 29
Lapata (1999c), 29
Lapata (2000), 29
Lapata (2001), 29
Lascarides and Copestake (1998), 19, 141, 231
Lascarides (1995), 19
Lauer (1995), 119, 175–179, 190, 195, 205, 207, 212, 215, 217, 232–234, 241
Leacock et al. (1998), 26
Leech et al. (1994), 32, 189
Lee (1999), 218
Leonard (1984), 174, 210, 212, 232, 233
Lesk (1986), 15, 23
Levin (1993), 3, 18, 20, 21, 24, 25, 27, 28, 49–53, 70–72, 74, 76, 77, 84, 86, 88, 90, 91, 94–96, 98, 99, 102, 104, 106, 107, 109, 111–113, 115–119, 122, 124–127, 131–137, 139, 168, 169, 173, 218, 238, 240
Levi (1978), 3, 18–20, 24, 25, 174, 194, 210, 211, 238
Li and Abe (1995), 107
Liberman and Sproat (1992), 174
Light (1996), 26, 27
Lin (1999), 181
Lodge (1981), 158
Lyons (1977), 140
Macleod et al. (1998), 33, 35, 154, 214
Manning and Schütze (1999), 26, 106, 175, 183
Manning (1993), 15, 26, 68, 104–106, 122
Marcus et al. (1993), 32
Marcus et al. (1994), 52, 99, 100
Marsh (1984), 174
Masterman (1957), 23
McCarthy and Korhonen (1998), 104, 107, 108
McCarthy (2000), 104, 108
McDonald (1982), 174, 212, 233
McDonald (2000), 244
Merlo and Stevenson (1999), 26, 104, 134, 206, 234
Mihalcea and Moldovan (1998), 171
Miller and Charles (1991), 57, 79, 133
Miller et al. (1990), 32, 33, 51, 66, 140, 182, 193, 216
Monsell (1991), 244
Mosteller and Wallace (1964), 213
Narayanan and Jurafsky (1998), 244
Ng and Lee (1996), 23
Ostler and Atkins (1992), 19
Palmer (2000), 113
Passonneau and Litman (1997), 206, 234
Patel et al. (1998), 133
Pedersen et al. (1996), 205
Pinker (1989), 50, 51
Pollard and Sag (1987), 60
Pollard and Sag (1994), 19, 73, 242
Poznański and Sanfilippo (1995), 15
Procter (1978), 23
Pustejovsky et al. (1993), 21, 57, 58
Pustejovsky (1995), 3, 19, 21, 24, 25, 140–142, 149–152, 154, 165, 169, 171, 193, 242
Quirk et al. (1985), 67, 68, 80, 140, 173, 214
Ratnaparkhi (1998), 61, 118
Redington et al. (1998), 244
Resnik and Diab (2000), 244

Resnik (1993), 15, 104, 107, 119, 135, 195, 206, 215–217, 244
Resnik (1997), 23
Roland and Jurafsky (2000), 103, 242
Rooth et al. (1999), 15
Russell and Norvig (1995), 38
Saffran et al. (1996), 244
Sapir (1944), 140
Schütze (1998), 15, 23
Schooler and Anderson (1997), 244
Schulte im Walde (1998), 104–106, 134, 135
Schulte im Walde (2000), 104, 134–136
Schütze (1993), 133
Schütze (1996), 46, 156
Selkirk (1982), 211
Shannon and Weaver (1949), 180
Siegel and Castellan (1988), 45
Siegel and McKeown (1994), 206, 234
Siegel (1999), 26, 206, 234
Smadja (1991), 15
Smadja (1992), 205
Small (1980), 22
Sparck Jones (1983), 176, 211
Spence and Owens (1990), 244
Spencer (1991), 61, 180
Spivey-Knowlton and Sedivy (1995), 66
Sproat (1994), 182
Stede (1998), 104, 107
Steedman (2000), 73
Stevenson and Merlo (2000), 45, 104, 134, 136
Stevens (1975), 46, 156
Strzalkowski and Vauthey (1992), 180
Sturt et al. (1999), 36
Sussna (1993), 23
Teufel (2000), 45
Vanderwende (1994), 212, 232, 233
Vendler (1968), 3, 19, 25, 140–142, 149–152, 154, 164, 168, 169, 171
Voorhees (1993), 23, 206
Warren (1978), 19, 175, 193, 194, 210, 211
Warren (1984), 19
Wechsler (1995), 60, 67
Weeber et al. (2000), 205
Weiss and Kulikowski (1991), 160
Whaley (1978), 244
Whittaker (1990), 43
Wiebe et al. (1998), 45
Wilks and Stevenson (1998), 23
Wilks et al. (1996), 23
Wilks (1975), 22
Wu (1993), 212, 233
Yarowsky (1992), 22, 206
Yarowsky (1993), 26
Yarowsky (1995), 15, 23, 132
Zamparelli (1998), 36