Lexical Acquisition at the Syntax-Semantics Interface: Diathesis Alternations, Subcategorization Frames and Selectional Preferences.

Diana McCarthy

Submitted for the degree of D.Phil.

University of Sussex

March, 2001

Declaration

I hereby declare that this thesis has not been submitted, either in the same or different form, to this or any other university for a degree.

Signature:

Acknowledgements

Firstly, I wish to thank Gerald Gazdar, my supervisor, for all his helpful comments, suggestions and words of encouragement throughout my doctorate. He has provided valuable feedback whenever I have requested it, and made painstaking efforts to steer me towards a thesis from the somewhat rough and ready drafts that I have presented him with. Secondly, a big thank you to John Carroll for his practical support and guidance whilst I worked with him on Sparkle, and throughout my thesis. I am indebted to him for suggesting this avenue of research. I am also grateful for the parses of the BNC that he provided. I thank him together with Ted Briscoe, for the use of their subcategorization frame acquisition system. Ted Briscoe has also provided me with invaluable feedback and suggestions on my research, whilst we worked on Sparkle, and long afterwards. I am also grateful to him for putting me in contact with Anna Korhonen. I have learned a great deal from discussions with Anna on subcategorization frame acquisition and diathesis alternations. I have also truly enjoyed collaborating with her on a couple of conference papers. I thank her for her efforts to produce a mapping between the subcategorization frame acquisition system and Levin's inventory of alternations, and for making this available to me.

I also wish to thank Stephen Clark for interesting discussions on selectional preferences, word sense disambiguation and structural disambiguation, and the arduous task of proof reading my thesis. I wish to thank Hang Li, Andreas Wagner, Ted Dunning, and Miles Osborne for discussions primarily on the minimum description length principle and selectional preference modelling. I have had varied stimulating discussions on word senses and their disambiguation with both Adam Kilgarriff and Mark Stevenson. Thanks also to Marc Light for an interesting discussion on the evaluation of selectional preferences and to Ted Pedersen for answering many queries that I had with regard to both the log-likelihood ratio and Fisher's exact test. I am grateful to David Weir for the loan of countless conference proceedings and to Bill Keller, who provided guidance alongside John and Gerald as part of my research committee. Thank you, again, to Gerald, John, Stephen and Bill for acting as human judges and providing a gold-standard for evaluation of the diathesis alternation identification approaches.

On a personal note, I wish to thank Martine Smets and Carol Shergold for their personal support and companionship. And to all others at COGS who helped in all sorts of ways when I was reduced to hobbling around the department on crutches. Finally, a big thank you to all my family for helping me to keep things in perspective, and for keeping their noses out!

Lexical Acquisition at the Syntax-Semantics Interface: Diathesis Alternations, Subcategorization Frames and Selectional Preferences.

Diana McCarthy

Summary

Verbs frequently permit their arguments to be expressed syntactically in more than one way. Such verbs are said to exhibit diathesis alternations. These alternations lie at the bridge between syntax and lexical semantics, since the semantics of a verb licenses its syntactic behaviour. This link between syntax and semantics is extremely useful for NLP purposes because it allows us to classify a verb semantically from its syntactic behaviour, and to predict further syntactic behaviour from the classification.

In this thesis, we propose a method for discovering verbs which participate in diathesis alternations where underlying arguments occur in different grammatical slots in the alternating variants. Rather than try to identify verbal participants by their semantic properties, our method exploits verbal argument structure and preferences which can be acquired automatically from corpora. We use acquired subcategorization frames to detect potential candidates for a given alternation. We then obtain selectional preference models for the grammatical slots, in the alternating frames, between which the arguments switch. We demonstrate a significant relationship between the similarity of the preference models, at the relevant grammatical slots, and verbal participation.

Although identification of diathesis alternation participants is the central goal of the thesis, we have two subgoals: the automatic acquisition of both subcategorization frames and selectional preference models. Since there is already a large body of research in these areas, we draw on the research of others. For subcategorization frame acquisition we use the original system of Briscoe & Carroll (1997). For the selectional preference models, we modify the approach devised by Li & Abe (Li & Abe, 1995; Abe & Li, 1996).

One of the main modifications we make for selectional preference acquisition, is to attempt word sense disambiguation of the argument heads used to create the preference models. We experiment with some methods which do not make excessive demands for training time or data.

Submitted for the degree of D.Phil.

University of Sussex

March, 2001

Contents

List of Acronyms and other Abbreviations

ACT 121
ANLT: Alvey Natural Language Tools . . . 13
αSD: α-skew divergence 133
ATCM: association tree cut model 32

BNC: British National Corpus 4

CAUS 121
CIDE: Cambridge Dictionary of International English 93
COMB 74

ED: euclidean distance 131
EM: expectation-maximisation 4

FN: false negative 52
FP: false positive 52
FREQ 67
FirstS 74

GATE: General Architecture for Text Engineering 4

HECTOR: 51
HMM: hidden Markov model 25
HPSG: Head-Driven Phrase Structure Grammar 112

IDN: ignore difficult nouns 68
INTR 121
SPass 74

LCS: lexical conceptual structures 112
LDOCE: Longman's Dictionary of Contemporary English 20
LLOCE: Longman Lexicon of Contemporary English 46
LLR: log-likelihood ratio 28
LLRTCM: LLR tree cut model 38
LO: lemma overlap 133
LOB: Lancaster-Oslo/Bergen Corpus 63

MCBL: multiple choice random baseline . 65
MDL: minimum description length 31
MRD: machine readable dictionary 2
MRT: machine readable thesaurus 54

NOWSD 74

PCP: probabilistic chart parser 136
POS: part of speech 2
PTCM: probabilistic tree cut model 32

RATIO 67
RBL: random baseline 65
RSA: role switching alternation 3

SCF: subcategorization frame 1

TCM: tree cut model 32
TN: true negative 52
TP: true positive 52

VBD 121

WSD: word sense disambiguation 6
WSJ: Wall Street Journal 12

List of Figures

List of Tables

Chapter 1

Introduction

This thesis concerns the automatic acquisition of verbal argument structure and selectional preferences. Verbs play a pivotal role in natural language as they are the key predicates in sentences, with all other constituents expressed in terms of their argument structure. There is a great deal of information that one might want to store in the verbal entries of a computational lexicon. Information about syntactic behaviour is crucial, since this information is central to successful parsing. Furthermore, the preferences that verbs have for the semantic type of their arguments has been used in many tasks, including structural disambiguation (Resnik, 1993b; Abe & Li, 1996) word sense disambiguation (Wilks & Stevenson, 1998b; Resnik, 1997; Federici, Montemagni, & Pirrelli, 1999), anaphora resolution (Ge, Hale, & Charniak, 1998; Murata, Isahara, & Nagao, 1999) and proper noun resolution (Wakao, Gaizauskas, & Wilks, 1996).

Verbs frequently permit their arguments to be expressed syntactically in more than one way. Such verbs are said to exhibit diathesis alternations. What is particularly interesting about these diathesis alternations is that they relate to both the syntactic behaviour and the lexical semantics of verbs. A particular alternation is typically associated with a number of verbs and the semantic properties of the participant verbs license the associated syntactic behaviour (Levin, 1993). In this thesis, we build on earlier work on the automatic acquisition of subcategorization frames (SCFs) and selectional preferences, and bring these two information sources together for automatically identifying verbal participation in diathesis alternations.

1.1 Automatic Lexical Acquisition

Automatic acquisition of lexical knowledge is an active area of research within NLP. This is because useful NLP systems will typically require lexicons with several thousand lexical entries, even for quite restricted domains. In a lexicalist approach, most information, such as subcategorization information, is pushed out of the grammar, and into the lexicon. This is arguably an appropriate place to express generalisations as well as idiosyncracies. Manually encoding all the lexical information required would be a costly enterprise, and certainly not a cost effective approach, since there are other ways of obtaining much of the information. Methods for automatic acquisition of lexical information have been developed for many areas, including collocations (Dunning, 1993;

Smadja, 1993), syntactic category (Finch & Chater, 1991; Schütze, 1993), word senses (Pereira, Tishby, & Lee, 1993; Schütze, 1992), subcategorization frames (Brent, 1991, 1993; Manning, 1993; Ushioda, Evans, Gibson, & Waibel, 1993; Briscoe & Carroll, 1997; Carroll & Rooth, 1998) and selectional preferences (Resnik, 1993a; Ribas, 1995a; Li & Abe, 1995; Pozanski & Sanfilippo, 1996; Abe & Li, 1996; Rooth, Riezler, Prescher, Carroll, & Beil, 1999). In this thesis we report our work on the acquisition of diathesis alternations, alongside other related work (Schulte im Walde, 1998; Lapata, 1999; Stevenson & Merlo, 1999).

Previously, lexical acquisition has been performed by obtaining data directly from machine readable dictionaries (MRDs) (Boguraev, Briscoe, Carroll, Carter, & Grover, 1987; Sanfilippo, 1994; Montemagni, 1994; Slator & Wilks, 1990), but this has been fraught with difficulties. MRDs are usually general purpose resources. They introduce many senses (and therefore ambiguities) not necessary, or relevant, to the domain and task at hand. They are built by human lexicographers for human readers and are therefore prone to human errors, inconsistencies and omissions (Briscoe & Carroll, 1997).

Most automatic acquisition is now done from corpora. Lexicons acquired from corpora are also subject to error, however the errors that they embody are of a different nature to those in human built resources. These errors arise from flaws in the software system processing the corpus, and also as a direct consequence of the use of naturally occurring data. Corpus data, even from a written source, is full of semantic anomalies and ungrammatical fragments of language. In automatic acquisition, these errors are considered as noise. A principal advantage of using corpus data is that frequency information is available. This is important in many NLP applications and crucial for statistical approaches. The frequency data, as well as the linguistic phenomena, is relevant to the corpus data from which it is acquired. The training data is selected to match the genre anticipated for the application. When changing to a new sublanguage it may be possible to simply retrain on material of the appropriate type. For radical differences in the corpus data, changes will probably be required to the software.

Although statistical lexical acquisition from corpora is now the norm, many researchers use some a priori knowledge to guide the collection of statistics (Gazdar, 1996; Klavans & Resnik, 1996). There are many ready made syntactic inventories, for example for SCFs or part of speech (POS) tags, which are suitable for use with many sublanguages. Existing semantic inventories, on the other hand, are usually hand-crafted for a particular sublanguage, or else general purpose resources are used. Hand-crafting a semantic resource for a particular sublanguage can require substantial up front effort. On the other hand, using a general purpose semantic resource brings with it the disadvantage that the knowledge contained therein is likely to be somewhat at odds with the text types required for the application. In either case, using a priori knowledge introduces the possibility of human error. The inadequacies of the a priori knowledge are, in principle, compensated for by the collection of corpus statistics over symbols supplied in advance. These are assumed to be appropriate, on the whole, for the task. Any human error introduced is diminished where it is not attested in the corpus data. The residual drawbacks are also compensated for by a potential increase in coverage (Li & Abe, 1996) and reduction in training time, compared to the use of automatically acquired classifications, which themselves fall prey to anomalous classes.

Li & Abe (1996) experimented with both human built semantic taxonomies and automatically

acquired classifications. They indicated different benefits associated with either approach. In the work reported in this thesis, we have endeavoured to acquire lexical information from corpora, but we have used knowledge from external sources to help structure this acquisition. We list the external sources below in section 1.6.

1.2 Diathesis Alternations

Diathesis alternations are different ways in which the arguments of a verb can be expressed syntactically. These alternations are typically accompanied by subtle changes in meaning and usually apply to a number of verbs. Diathesis alternations lie at the bridge between lexical semantics and syntax since there is a strong relationship between a verb's participation in specific alternations and the semantic properties of the verb. Levin (1993) has manually produced a semantic classification of over 3000 verbs based on their participation in her inventory of 80 diathesis alternations. This inventory deals mainly with alternations involving noun phrase (NP) and prepositional phrase (PP) constituents. Examples of alternations include:

The dative:

(1) a. She gave the dog a bone.

b. She gave a bone to the dog.

The causative-inchoative alternation:

(2) a. The boy broke the window.

b. The window broke.

The implicit object construction:

(3) a. The boy ate the popcorn.

b. The boy ate.

In the first two alternations, arguments of a particular semantic type have different grammatical relationships with the verb in the alternate syntactic realizations. In the second and third alternations, an argument is omitted. In this thesis, we propose a method to automatically acquire alternations of the first type, whether or not an argument is omitted in one of the alternating variants. We refer to these as 'role switching alternations' (RSAs).

1.3 Our Contribution

We propose a method to directly detect the switch of a particular argument type between different grammatical slots in alternating frames. We use SCF information to identify verbal candidates with the relevant syntactic behaviour, and to obtain argument heads specific to the appropriate slot and SCF. We then use selectional preference information to find cases where the semantic type of the argument in one grammatical slot does in fact switch position to another grammatical slot in the alternating variant.

The SCF and selectional preference information that we use are acquired automatically. SCF acquisition and selectional preference acquisition therefore present subgoals for our thesis. However, a considerable amount of research has been done in these areas and so we use this research where appropriate, with modifications where necessary.

1.3.1 SCF Acquisition

There have been a substantial number of contributions in this field (Brent, 1991, 1993; Manning, 1993; Ushioda et al., 1993; Briscoe & Carroll, 1997; Rooth et al., 1999; Korhonen, Gorrell, & McCarthy, 2000). Brent (1991) showed that a set of 5 SCFs could be recognised successfully using information from unambiguous cases. He avoided the need for syntactic analysis by, for example, using only pronouns to detect noun phrases. He used a binomial hypothesis test to statistically filter out cases where the cue had occurred less than would be expected by chance. Since he only used unambiguous cases, he could not provide frequency information, or even a rank order, alongside the SCF information. Such information is valuable for any computational lexicon, and particularly necessary for any statistical model making use of the SCF entries. For this reason, other researchers in this field have sought evidence from all examples from the training data, and this has necessitated syntactic analysis of the training data.

Ushioda et al. (1993) and Manning (1993) both used POS tagged data and finite state NP parsers. Ushioda's system recognised six SCFs and provided frequency information alongside this. Manning used syntactic information, like Ushioda, but also used a statistical filter, like Brent. He permitted less reliable cues than Brent and relied on a stricter mechanism of filtering to ensure reliability. He recognised 19 SCFs, though some of these were parameterised by a preposition.

Briscoe & Carroll (1997) developed a more comprehensive SCF acquisition system. This system distinguishes 161 SCFs, and returns relative frequencies for each SCF found for a given verb. The input text is POS tagged using an HMM tagger (Elworthy, 1994) and the CLAWS-2 tagset (Garside, Leech, & Sampson, 1987). The tagged text is then lemmatised with an enhanced version of the morphological analyser provided in the General Architecture for Text Engineering (GATE) (Cunningham, Gaizauskas, & Wilks, 1995) software environment. The lemmatised text is fed to a shallow parser. A patternset extractor operates on the parser output and produces subcategorization patterns from the shallow parses. These patterns are classified according to the inventory of 161 SCF classes, or rejected as unclassifiable. Finally, the hypothesised SCF entries are filtered to remove those for which the quantity of evidence is less than or equal to that expected by chance.

Carroll & Rooth (1998) proposed an iterative approach for estimating SCFs. They used a probabilistic version of a manually developed context-free grammar of English. They trained this using the expectation-maximisation (EM) algorithm, and lexicalised the grammar with argument heads detected using the grammar rules. They ran the EM algorithm again to estimate the expected frequencies of a head word occurring with specified SCFs. Probability estimates were then fed back into the grammar for the next iteration. Carroll & Rooth reported promising results for three verbs on applying their technique to the British National Corpus (BNC) (Leech, 1992).

The performance in terms of precision and recall varies depending on the test data, but many systems achieve somewhere around 80% token recall. This is particularly impressive for the

Briscoe & Carroll (1997) system since they have to differentiate between 161 SCF types. We use this acquisition system for the work described in this thesis. The large inventory of SCFs that the system handles is helpful for finding candidates with appropriate syntactic behaviour for identifying diathesis alternations. Additionally, argument head data is supplied at specified slots at the entries for verb and SCF. We do not propose any further modifications to this SCF acquisition system in this thesis. [1]

We use the SCF lexicon from the SCF acquisition system for identifying RSAs. The SCF acquisition system is used to establish candidate verbs which take the alternating SCFs involved in the RSA. To specify the SCFs involved in a target alternation, we utilise a mapping between the SCFs of the Briscoe & Carroll SCF acquisition system and the alternations specified in Levin. [2] We hereafter refer to this mapping as the Levin–SCF mapping.

For many alternations, information from the SCF system is not, by itself, sufficient for diathesis alternation identification. However, the SCF lexicon can be used directly in cases where it is. Generally speaking, evidence from the argument heads is required and in these cases the entries in the SCF lexicon, which provide argument head data, can be fed to our selectional preference acquisition system. Tuples of the form <verb, slot, lemma>, for example $< eat, direct\ object, biscuit >$, are created from the SCF entries. In these tuples, the slot relates to the specific syntactic slot of the verbal predicate, and for diathesis alternation identification this also specifies the SCF frame in which it appears. The lemma is an argument head which occurred in the training data at this specified slot. In this thesis we only deal with arguments which are expressed as noun phrases (NPs) or prepositional phrases (PPs). The argument head in both cases is the noun heading the NP. For PPs, the preposition is specified along with the verb in the tuple: <verb:prep, slot, lemma>. For diathesis alternation identification, the slots of the SCFs at which the role switching occurs will be referred to hereafter as the target slots. For example, the target slots of the causative alternation are the direct object slot of the transitive SCF, and the subject slot of the intransitive SCF.

1.4 Selectional Preference Acquisition

In chapter 2, we provide a full account of the background to this area, and describe the system which we modify. A large variety of systems have been already proposed. These can be broadly categorized as lemma based systems, class-based systems or similarity-based systems. Lemma based systems use the argument heads directly. Class-based systems reduce the problems of sparse data by permitting generalisations where a specific <verb, slot, lemma> has not been seen. In similarity-based systems, distributional evidence is used for smoothing, without explicitly producing classes. For diathesis alternation identification, we advocate the use of class-based selectional preferences to reduce the problems caused by sparse data: many lemmas in one target slot will not also occur at the target slot in the alternating SCF in any naturally occurring sample of corpus data, even though they are acceptable in both slots. However, we use a lemma-based experiment as a baseline in chapter 5, for comparison with class-based preference models.

The modifications that we have applied to the selectional preference acquisition method in-

[1] Although there is scope for further modifications and we have collaborated with other researchers working in this area (Korhonen et al., 2000).

[2] The mapping is the work of Anna Korhonen. We are indebted to her for the use of it.

clude the use of word sense disambiguation (WSD) on the argument heads and the classification of proper nouns. WSD is a vast area of research, and we have investigated several possibilities for disambiguation of the argument heads.

1.5 Diathesis Alternation Identification

There has been some recent interest in the automatic acquisition of diathesis alternations (Resnik, 1993a; McCarthy & Korhonen, 1998; Schulte im Walde, 1998; Lapata, 1999; Stevenson & Merlo, 1999; Rooth et al., 1999; McCarthy, 2000). The early work by Resnik was specific to implicit object alternations, characterised by the transitive and intransitive frames, where the direct object of the transitive frame is omitted in the intransitive. This contrasts with the causative-inchoative alternation which involves what we are calling a role switch, where the object of the transitive becomes the subject of the intransitive frame. Resnik used his measure of selectional preference to test the theory that participation depends on the ease with which the omitted object is inferred.

More recently there has been some interest in identifying alternations generally. The automatic acquisition of diathesis alternations has been made possible by technological advances in robust parsing which have lead to the acquisition of syntactic information from corpora. In addition to syntactic information, semantic information is required in many cases. The semantic category of the verb would certainly help establish participation, used together with a means of mapping to Levin's classification (Dorr & Jones, 1996). However, this approach will not help in cases where the semantic class of the verb is unknown, and it increases our reliance on manmade resources, where these are used to define the semantic categories. Schulte im Walde (1998) clustered automatically acquired SCF information and demonstrated a significant overlap between the verb clusters and the verbs in Levin's classification. This overlap, measured in terms of precision and recall, was reduced when she added automatically acquired preferences as features for clustering. Lapata (1999) used manually determined semantic cues on the argument slots. Her approach necessitated the use of a priori semantic knowledge specific to the alternation. For some alternations this may be straightforward, for others the semantics may be harder to stipulate. Another approach is to use cues for syntactic frames, coupled with the overlap of lexical fillers between the alternating slots (Stevenson & Merlo, 1999). Again, the features used to distinguish the alternation behaviour are specified a priori and are specific to the distinctions being made. Furthermore, in many cases, lexical overlap will not be a reliable indicator because of the sparseness of the data.

1.5.1 System Overview

In this thesis we propose a method for identifying RSAs which is generally applicable and which does not require hand-coded knowledge specific to the alternations. The SCF acquisition system and the Levin–SCF mapping are used to identify potential candidates for the RSAs. The selectional preference acquisition system is used on the argument head data, which is stored at the slots of the SCF lexicon specified by the mapping. WSD is optionally applied to the argument head data in the SCF lexicon. The selectional preference models can themselves be used for WSD in an iterative approach. The major system components are shown in figure 1.1.

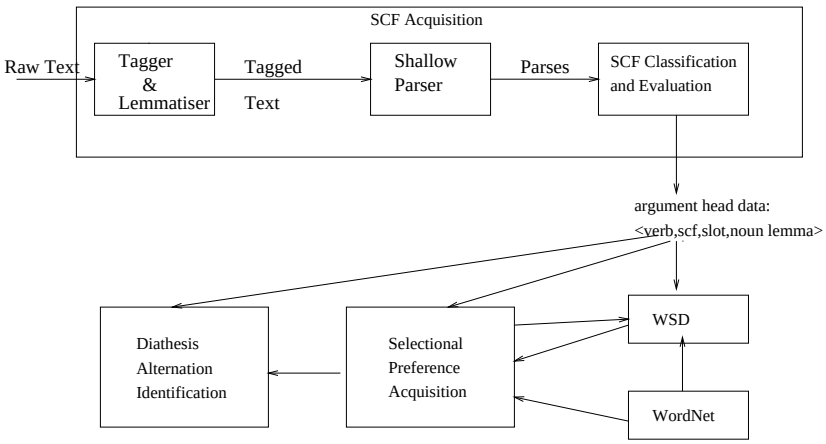

Figure 1.1: System overview

1.6 External Resources Used

1.6.1 Software

We used the Briscoe & Carroll SCF acquisition system with two different parsers. Initially we used text parsed with a probabilistic chart parser (Chitrao & Grishman, 1990). Subsequent work used SCF lexicons built from the output of an LR parser (Inui, Sornlertlamvanich, Tanaka, & Tokunaga, 1997). The performance of the two parsers is compared in chapter 5; the differences were not found to be statistically significant.

For named entity recognition we used the General Architecture for Text Engineering (GATE) (Cunningham et al., 1995) software environment. This includes a named entity recognition component. We used the VIE NE recognition system in GATE version 1.1.

For our selectional preference models, we used the noun hyponym hierarchy in WordNet (Beckwith, Fellbaum, Gross, & Miller, 1991; Miller, Beckwith, Fellbaum, Gross, & Miller, 1993b; Fellbaum, 1998), version 1.5. We collected corpus statistics for the classes in this hierarchy. WordNet has a wide coverage and is freely available, however, we acknowledge that use of this man-made resource constrains our application to the structure of the taxonomy, and the sense inventory defined within it. Alongside WordNet, we used the SemCor corpus (Miller, Leacock, Tengi, & Bunker, 1993a). SemCor is a 250,000 word portion of the Brown corpus (Francis & Kučera, 1979) that has been manually tagged with WordNet sense tags. It is freely available along with WordNet. We used this corpus to obtain sense frequency information and as a resource for WSD evaluation.

We used the written part of the BNC for our lexical acquisition experiments. The written portion totals 90 million words. We worked with parsed portions of this as they were made available to us. However, because our methods are automatic we could process further portions in a similar manner, given sufficient processing time. We used 4 different SCF lexicons acquired from portions of this corpus:

1. Lexicon A - acquired from 10.8 million words of parsed text
2. Lexicon B - acquired from 1.8 million words of parsed text

3. Lexicon C - acquired from 1.8 million words of parsed text, with proper noun recognition
4. Lexicon D - acquired from 19.3 million words of parsed text

The lexicons A, B, and C were all parsed with the probabilistic chart parser, whilst lexicon D was parsed with the LR parser. The portion of the BNC used for lexicon A was a subpart of the sample parsed for lexicon D. The lexicons B and C were both produced from the same portion of the BNC, but proper noun recognition was performed on the data for lexicon B, whilst proper nouns were left unclassified in the other lexicons. The data for the lexicons B and C was a subpart of the sample parsed for lexicon A.

1.7 Chapter Summaries

Chapter 2. This chapter describes what has already been achieved in the area of selectional preference acquisition. We describe the approach that we adopted, originally devised by Li & Abe (Li & Abe, 1995; Abe & Li, 1996), and our reasons for adopting it. We describe the modifications that we made to this approach, along with our rationale for these modifications.

Chapter 3. This chapter tackles issues concerned with sense tagging argument head data for input to the preference acquisition system. We survey some WSD approaches and select three which we took to be most relevant for our task. We experimented with these three approaches to identify the levels of performance, and the resources required for training and running the WSD modules. One approach was not taken forward for selectional preference acquisition because of performance problems with randomly selected nouns, as opposed to nouns with clear sense distinctions, and because of the computational cost required for training. We then applied two WSD modules to the argument head data, and informally compared the differences that these make. The first approach was a simple first sense heuristic. The second was an iterative approach, using the selectional preferences themselves for WSD of the argument head data. We also experimented with a combination of these two modules.

Chapter 4. In this chapter, we provide details of our formal evaluation of the selectional preference acquisition system. We describe general strategies for the evaluation of automatically acquired lexical acquisition, and then describe what has already been done in the area of selectional preference acquisition in terms of these strategies. We selected evaluation methods for two reasons: (i) to compare our results with the results of other researchers in this area, and (ii) to compare the different parameter options within our selectional preference acquisition system.

Chapter 5. This chapter presents the main contribution of the thesis. Here we bring together the SCF and selectional preference acquisition systems for the automatic identification of diathesis alternations. For two RSAs, subcategorization information alone was sufficient to determine participants. For two other RSAs, we present the results of two different approaches using selectional preferences. The first approach was rejected since it only worked in cases where the alternating frames have a similar frequency of occurrence. The second approach worked with both RSAs for which we had sufficient test data. These results are contrasted with those obtained using a measure of lemma overlap. Unlike the class-based approach, this approach did not reliably show a significant relationship between the similarity measure and verbal participation in the target alternation, where verbal participation was determined by human judges.

Chapter 6. We finish with a summary of the contribution of the thesis, and directions for future research.

Chapter 2

Selectional Preference Acquisition

Selectional preferences are the semantic tendencies that predicates have for their arguments. The concept of 'selectional' or 'semantic preference' arises from 'Preference Semantics' (Wilks, 1975b, 1975a). 'Preference Semantics' is an approach to understanding natural language utterances in terms of the semantic requirements of the words for the context that they can occur in. Wilks's notion of semantic requirements contrasts with Katz and Fodor style restrictions (Katz & Fodor, 1964). Restrictions in the Katz and Fodor tradition provide hard and fast constraints that amount to violations when they are not met. Preferences on the other hand imply a gradation whereby alternative analyses can be ranked. In Wilks's scheme, word combinations are analysed in terms of a core set of primitive semantic units which can be combined in preset ways. When analysing an utterance the constraints can be broken and no analysis is ruled out. The analysis that is preferred satisfies more of these constraints than the other analyses, thereby providing the greatest 'semantic density' for the utterance. Researchers in NLP have adopted preferences more readily than restrictions. Finding an exclusive set of lemmas, lemma classes or features to represent restrictions is not appealing since many items appear in the data in places where our intuition tells us they should not. Preferences allow us more readily to cope with real data. Moreover preferences acquired automatically from empirical data allow us to avoid the difficult laborious introspections encountered when devising the preferences manually.

In this chapter we outline current methods for acquiring preferences between predicates and arguments. We adopt one of these methods for this thesis and at the end of the chapter we describe some alterations that we made to the original approach.

2.1 Uses of Selectional Preferences

Preferences are frequently used for processing naturally occurring data. They capture the lexical information which is often sought for resolving both structural and lexical ambiguity. The preferences that hold between verbs and the argument head of prepositional phrases have been exploited as a means of resolving prepositional phrase ambiguity (Resnik, 1993a, 1993b; Li & Abe, 1995; Abe & Li, 1996). This involves a situation where the prepositional phrase is potentially a modifier of the NP or is acting as an adjunct or argument of the verb, more typically an adjunct. For example

in analysing the sentence:

(4) He hit the man with the stick.

It is important to determine whether the prepositional phrase should be attached to the verb *hit* or to *the man*. This will determine the appropriate interpretation i.e. whether the stick is used for hitting the man or whether the man who is hit has the stick.

Preferences have also been used for disambiguating word senses (Resnik, 1997; Ribas, 1995a). In example 5(a) below we would expect the **financial institution** sense of *bank* to be more likely than the **raised strip of earth** sense given the verb *rob*. However, in the second example one would expect the latter reading.

(5) a. She robbed the bank.

b. She slid down the bank.

Anaphora is another field where preferences have been applied (Ge et al., 1998). Two contrasting sentences exemplify how these might help:

(6) a. The packet contained chocolate but nobody was allowed to open it.

b. The packet contained chocolate but nobody was allowed to eat it.

In 6(a), the referent of *it* is clearly the *packet* since this is more likely as an object of *open*, meanwhile in the 6(b) *chocolate* would be more likely as the object of *eat*. Such a task has not yet been performed using automatically acquired preferences because of the lack of a suitably annotated corpus. Weeding out the cases of anaphora suitable for application of preferences and marking up potential referents is a good deal more complex than obtaining target instances for prepositional phrase disambiguation or word sense disambiguation.

In recent years, there has been a growing trend for the speech community to look into incorporating selectional preferences into speech understanding. A field which has traditionally used statistical pattern matching in the form of hidden Markov models and neural networks (Price, 1996) The nature and representation of the selectional preferences will of course vary but the intuition is that preferences can help choose between alternate arguments given that we have determined the predicate.

The application should have some bearing on the nature of the representation of the selectional preferences. In addition to their use for natural language processing, selectional preferences are of use to lexicographers (Ribas, 1995a). In this case, the end-product of preference acquisition must be in a form which is easily read by humans.

The ultimate goal of this thesis is to establish a method in which selectional preference models can be used as a means of identifying verbs where the underlying arguments (such as agent or patient) can appear in different grammatical roles (such as subject or direct object) in different SCFs. In our terminology, the selectional preference models can be used to establish verbal participation in RSAs. The acquisition of selectional preferences models is a prerequisite, however, the acquisition of selectional preference models is itself an important by-product of this research.

There is already a substantial body of research on automatic acquisition of preferences. Rather than duplicate these efforts, we select from the approaches provided by other researchers, who

are typically using preferences for structural and lexical disambiguation. We then supplement the acquisition process with some modifications to the core technique. These are described in this chapter. In the next chapter, we experiment with some WSD techniques for coping with the massive ambiguity of the input data.

2.2 Background

A decade ago, selectional preferences or restrictions were produced manually from introspection. Nowadays, such an approach would justly be shunned because manual endeavours would be overwhelmed by the quantity of data needed. The strong appeal of an automatic approach is the avoidance of labour intensive methods. Moreover, reliance on human intuition falls preys to the human errors that beset any substantial manmade resource of this nature.

In the following sections, we give an account of the automatic approaches used to acquire preferences. Rather than compare the approaches one by one in sequence, we compare them in terms of the data used (see section 2.2.1) and the representation (section 2.2.2). We then go into more detail on the systems closest to our own in section 2.3. In section 2.3.3 we give a closer account of the approach we adopt, originally proposed by Li & Abe, and in section 2.4 we describe the changes we make to the system provided by Li & Abe.

2.2.1 The Data

Typically automatic approaches involve extracting predicate argument relationship tuples from machine readable dictionaries or corpora. The majority view favours corpora. Dictionaries bring with them a substantial amount of up-front effort from the lexicographers which at first glance is appealing. For example, if we collect typical argument heads as Montemagni (1994) did, then these will be stored by sense of the verb rather than by verb form. Of course, relying on human preprocessing of the data in this way leaves us open to the manmade errors we are trying to avoid when using automatic acquisition. A powerful draw of using corpora is that we can acquire information specific to the domain of the corpus. Portability to another domain is then just a matter of having an appropriate corpus. A further significant drawback of using dictionaries is that they do not possess the frequency information that naturally occurs within corpora. This is important when acquiring preferences along a continuum, rather than restrictions, since some measure of preference is required and frequency counts from naturally occurring data provide an obvious source.

Many researchers (Ribas, 1995a; Resnik, 1993a; Li & Abe, 1995; Abe & Li, 1996) needing corpus data for this purpose have used the Penn Tree Bank II (Marcus et al., 1995). This corpus provides a useful source of 2.6 million words of parsed text (1 million of Wall Street Journal (WSJ) articles, 1 million from the Brown corpus and the rest from other sources). The parsers have been produced by the Fidditch parser (a deterministic parser) with hand-correction of the output. Although this corpus is a useful source of syntactic relations between lexical items there is the distinct disadvantage that for new data the manual correction involved would again be needed. For many verb and slot combinations the Penn Tree Bank may simply not hold the quantity of data required. This is particularly important because we ultimately wish to look at diathesis alternations. Specific subcategorization frames (SCF) are less frequent than general slots, such as the direct ob-

ject, which occur within a multitude of frames. Alternations between these frames are rarer still. We require a data source where, given sufficient processing time, we can process as much data as we want without substantial human effort. Other researchers (Grishman & Sterling, 1993; Abney & Light, 1999) allow for this by using the output from fully automatic parsers to produce the tuple data.

The work described here employs a SCF lexicon built automatically from parses produced by a fully automatic shallow parser. The system that produces this is described by Briscoe & Carroll (1997). There are three main advantages to using this as the start point instead of the output from the parser.

Firstly, the shallow parses are classified according to preconceived subcategorization patterns. There are 161 subcategorization classes in all, giving a fine level of granularity of the frames. The 161 classes are a superset of those found in the Alvey Natural Language tools (ANLT) dictionary (Boguraev et al., 1987) and the COMLEX Syntax dictionary (Grishman, Macleod, & Meyers, 1994). The lexicon is organised by verb form with sub-entries for each SCF the verb participates in. The argument head tokens found in the corpus are listed at the appropriate slot within a SCF entry for a verb. For example, the entry for the transitive class for *bake* in a lexicon produced from 10.8 million words of parsed text from the BNC (lexicon A) is shown in figure 2.1. This figure displays the entry as it was output from the SCF acquisition system. The data that we are interested in is the :TARGET, which specifies the verb, the *:CLASSES* which specifies the SCF classification and the lists of lemmas with POS tags at :SLTL and :OLT1L which are the argument heads appearing at the subject and first argument position respectively, of the specified SCF. Each SCF classification is represented by a list of one or more SCF class numbers, 24, 51 and 161 in the example in figure 2.1. More than one SCF class is provided by the SCF acquisition system where, for some subcategorization patterns, the system cannot tell which of the classes is appropriate, so all the possible classes are provided. In this case, the system cannot distinguish different control options. Class 24 is used for plain transitive frames, exemplified in 7(a). Class 51 is intended for raising verbs, such as *seem* in 7(b) and class 161 is intended for equi verbs, such as *feel* in 7(c).

(7) a. He loved peas.

b. He seemed a fool.

c. He felt a fool.

The subcategorization lexicon furnishes us with the means to collect data not only specific to a particular slot, as other researchers do, but also with the option of being specific to the SCF. When acquiring selectional preferences for general disambiguation purposes, one might not want to go into this much detail. But, for our ultimate goal of observing preferences in slots of specific frames for diathesis participation identification, this is crucial.

The second advantage of using the lexicon is that after initial detection of subcategorization patterns, a statistical filter is applied to determine whether verbs co-occurring with frames do so with sufficient evidence for this combination not to have arisen by chance. The filter is designed to remove instances where there is insufficient evidence for the verb, given the evidence for the frame irrespective of verb. This filter is designed to remove some of the noise from the data, which

Figure 2.1: SCF lexicon entry for *bake* transitive class

```
Legend:
#S(EPATTERN  :TARGET |<verb>| :SUBCAT (VSUBCAT <syntax of detected arguments>)
:CLASSES ((<SCF classification>) <frequency of the SCF classes in the ANLT dictionary>)
:RELIABILITY <parser reliability>
:FREQSCORE <value assigned by the statistical filter>  :FREQCNT <number of tokens>
:TLTL  (<POS tags for the verbs in  each token>)
:SLTL (<argument heads in subject position for each token, with POS tag>)
:OLT1L (<argument heads in 1st argument position for each token, with POS tag>)
:OLT2L (<argument heads in 2nd argument position for each token, with POS tag>)
:OLT3L (<argument heads in 3rd argument position for each token, with POS tag>)
:LRL <indicates frequency of any lexical rules applied, e.g. passive>)



 #S(EPATTERN :TARGET |bake| :SUBCAT (VSUBCAT NP)
             :CLASSES ((24 51 161) 5293) :RELIABILITY 1.0
             :FREQSCORE 0.0 :FREQCNT 30
             :TLTL
             (VVG VV0 VV0 VV0 VVD VV0 VV0 VV0 VV0 VV0 VV0 VVG VVG VVG
              VVD VV0 VV0 VV0 VV0 VV0 VV0 VV0 VV0 VV0 VV0 VV0 VV0 VV0
              VV0 VV0)
             :SLTL
             ((|she| PPHS1)) ((|woman| NN1)) ((|i| PN1))
              ((|they| PPHS2)) ((|you| PPY)) ((|you| PPY))
              ((|society| NN)) ((|teaspoon| NN2)) ((|it| PPH1))
              ((|mother-in-law| NN1)) ((|you| PPY)) ((|you| PPY))
              ((|they| PPHS2)) ((|she| PPHS1)) ((|layer| NN1))
              ((|it| PPH1)) ((|it| PPH1)) ((|bit| NN2)) ((|David| NP))
              ((I PPIS1)))
             :OLT1L
             ((|scone| NN2)) ((|cake| NN2)) ((|cake| NN1))
              ((|them| PPHO2)) ((|bread| NN1)) ((|cake| NN1))
              ((|anything| PN1)) ((|cake| NN2)) ((|potato| NN2))
              ((|potato| NN2)) ((|cake| NN1)) ((|egg| NN1))
              ((|cake| NN1)) ((|clicking| NN1) (|bread| NN1))
              ((|cake| NN1)) ((|pipe| NN2)) ((|cake| NN1))
              ((|food| NN1)) ((|that| DD1)) ((|sheet| NN2))
              ((|it| PPH1)) ((|batch| NN1)) ((|foot| NN2))
              ((|cake| NN1)) ((|scone| NN2)) ((|rock| NN1))
              ((|potato| NN2)) ((|potato| NN2)) ((|pie| NN1))
              (|bread| NN1))
             :OLT2L NIL :OLT3L NIL :LRL 0)
```

does not receive any manual correction. Particularly this filter is aimed at removing adjuncts on the premise that they do not occur with individual verbs more than they would do so by chance.

Thirdly, SCF information is also valuable when producing preferences for PPs. There we can use the classification to determine genuine PPs as opposed to cases where the preposition is really acting as a multi-word complementizer. For example, *he seems as if he is clever* is marked by the grammar as having a preposition (or particle) but the subcategorization pattern extraction stage correctly identifies this as having a sentential complement.

Research has concentrated on slots involving NPs (subjects and direct objects) and PPs. This is presumably because there is a clear relationship between the argument head of the NPs in these slots and the verbal predicate. This thesis also concentrates on these slots.

2.2.2 Representation

The choice of representation has been the basis on which other researchers have classified alternate approaches (Resnik, 1993a; Ribas, 1995b). They characterise a three-way split between using words, automatic classifications from distributional evidence, and manually created word-classes. The choice of representation is bound up with the preference extraction process. For some of the systems mentioned here, there is more to say on the actual acquisition process and this will be done in section 2.3.

Acquisition of selectional preferences is usually performed by examining the contexts of the predicates (usually verbs) where the contexts are the head arguments in specified slots (usually nouns in subject, object and PP slots). There are, however, techniques originally proposed for extracting other relationships which are relevant since they can be readily transferred to the predicate argument relationship. For example, Church et al. (1991) investigated adjacent words, whilst McKeown & Hatzivassiloglou (1993b) examined the relationship between adjectives and the nouns they modify. To a large extent, these methods are transferable. Schütze (1992) avoided pre-processing of the raw text data by using 'window contexts' where the window was demarcated using a fixed distance from the target word. The choice of the representation is, to some extent, independent of the relationship being investigated. We thus overview techniques that have been used for other relationships since they could be applied to the arguments of verbal predicates.

Using Words

Some researchers have experimented using word forms themselves as the basis of capturing the relationship held between predicate and argument (Church et al., 1991; Hindle & Rooth, 1991, 1993) The method used is to take a given relationship and calculate statistics such as mutual information or the t-test between the predicate and the words in the slot under examination.

There are two main problems with this approach. Firstly, word based methods reflect word forms rather than senses and it is the latter which are relevant for semantic constraints. For structural disambiguation tasks this may not prove too much of a problem since the word forms may be informative enough to rank analyses. For some lexical disambiguation tasks this may present problems. Word forms associated with specific verbs and slots could not by themselves be used for WSD of fresh argument head data. Secondly, using word forms leaves us unable to generalise to new data unless we apply some sort of smoothing. Word-based techniques without smoothing are probably better suited to discovering idiosyncratic collocations (Smadja, 1993), where we would

not wish to make generalisations to word classes. Since we are interested in preferences which will cope with novel combinations, we do not dwell on this approach but turn to approaches that allow some means of generalising by smoothing the frequency distribution using distributional evidence.

Using Distributional Evidence

These approaches all exploit the tendency for words with similar semantics to occur in the same sorts of contexts. They accord with Firth's observation:

> You shall know a word by the company it keeps! (Firth, 1957, pg.11)

In these techniques, the distributional evidence of words which occur in similar contexts is used for generalizing and smoothing. The words can be clustered to provide automatically produced classifications using the distributional evidence (Schütze, 1992; Schütze & Pederson, 1995; Schütze, 1998; Pereira et al., 1993; McKeown & Hatzivassiloglou, 1993a, 1993b; Rooth et al., 1999), the proximity measures can be used explicitly for smoothing without actually producing a classification (Grishman & Sterling, 1993), or the whole data set can be stored as an 'example-base' which can be used to compare analyses for novel combinations (Federici, Montemagni, & Pirrelli, 1997; Federici et al., 1999).

Of these three variants, clustering is the most popular. Automatic classifications have been produced for a variety of different relationships and applications. Different clustering algorithms and distance measures have been used to compare the context vectors, however, the fundamental principle is to cluster words according to the distribution of the other words with which they co-occur in the specified relationship. A great many clustering techniques and similarity measures have been experimented with, and we do not attempt to describe or list them all here. We will instead outline four as examples.

McKeown & Hatzivassiloglou (1993b) clustered adjectives with reference to the distributions of nouns which they were found to modify. The distributions were compared using Kendal's τ coefficient. In this study, linguistic knowledge was used in addition to the occurrence data. If two adjectives appeared together, modifying a noun, then this provided strong negative evidence against these two adjectives being grouped together, on the basis that each adjective must be adding something new. For example *the old rusty car* would indicate that *old* and *rusty* should not belong to the same class.

In the work of Pereira et al. (1993), nouns were represented using the probability distribution of the co-occurring verbs with which the nouns appeared as direct objects. Relative entropy was used as a distance measure between these distributions. The clustering process was hierarchical and clusters were formed on the basis that they preserve entropy (information) as much as possible. Membership of the clusters was probabilistic so that words belonged to more than one class, and membership was a matter of degree. Clustering was performed using an expectation-maximisation (EM) algorithm (Dempster, Laird, & Rubin, 1977)

Rooth et al. (1999) also performed clustering using the EM algorithm. Unlike other clustering approaches, the classes created contained both argument heads and verbal predicates, moreover the verbal predicates were specified along with a slot and SCF. For example, a class was created

	witness	suit	laundry
court	320	240	80
clothes	30	280	300

Table 2.1: A collocation matrix

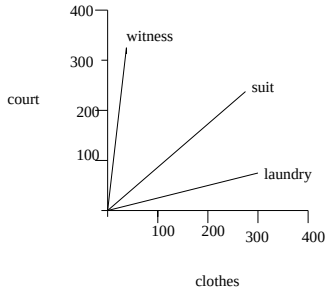

Figure 2.2: Schütze's semantic space in two dimensions

with verb and slot combinations including:

increase:subj:intrans, *increase:obj:trans*, *fall:subj:intrans*, *decline:subj:intrans*

and nouns including:

number, rate, price, amount

The class was interpreted as one involving verbs denoting scalar change and nouns denoting things which can move along scales. In this system, the classes smoothed the data so that the verbs (with SCF and slot, vs) and nouns (n) were not conditioned directly on each other, but on the classes ($c \in C$):

$$p(vs,n) = \sum_{c \in C} p(c,vs,n) = \sum_{c \in C} p(c)p(vs|c)p(n|c) \tag{2.1}$$

The EM algorithm was used to find classes which provided a joint distribution for verb-noun pairs which accorded well with the observed data.

Schütze (1992) avoided any pre-processing of the raw text data by using 'window contexts' a fixed distance from the target word. He used a character based window size rather than word based because longer words provide more information than shorter words, which are more likely to be closed class function words. He collected collocation data from within the window contexts and then clustered words according to these co-occurrences. Thus for example, given a simple two dimensional semantic space defined by only the context words *clothes* and *court*, the co-occurrence matrix given in table 2.1 could be pictured graphically as in figure 2.2.

Ambiguous words, such as *suit* were positioned somewhere between the positions of their respective senses, biased towards the more frequently occurring senses. These context vectors

were then clustered automatically using the cosine between the vectors as the similarity metric. Computationally this was expensive both in time and space because Schütze used large windows of up to 1200 words, and the corpus had more than 50 million words, so there were few cases of a word not occurring at all within the window of another word and hence there were not many zero's in the collocation matrices. To get over this, he used dimensionality reduction by means of single value decomposition to produce a more efficient distributed representation. Singular value decomposition selects axes in *n*-dimensional space which have the largest variation for the items being clustered. In this case, *n* is the number of word types used to represent semantic space. In addition to providing a more efficient representation, the dimensionality reduction achieves smoothing by reducing the noise in the original data.

The chief disadvantage of clustering on the basis of distributional information is that the words in the classes produced are not always semantically similar. Moreover, sometimes there is no obvious relation between the class members. For example, Pereira et al. reported a cluster including the words *pollution, increase* and *failure*, also one including *state*, *modern* and *farmer*. As Resnik describes it

> It would seem that the information captured using these techniques is not precisely syntactic nor purely semantic — in some sense the only word that appears to fit is *distributional* (Resnik, 1993a, pg.18)

In order to get a coherent set of classes, some researchers turn to manual editing after automatic classification has taken place (Sekine, Ananiadou, Carroll, & Tsuji, 1992). Others (Basili, Pazienza, & Velardi, 1993) spend human effort on semantically categorising the input text before clustering takes place.

Lexical ambiguity, which affects all representations of preference, is an additional problem. Evidence collected from corpora collapses word senses into word forms since it is the latter that are observed. In the work of both Pereira et al. (1993) and Rooth et al. (1999), some allowance was made for this as the clusters were 'soft' rather than hard Boolean ones. Membership was probabilistic and so a word can belong to more than one cluster. However, because the data from different senses was combined, a word type will still be positioned within clusters somewhere between the places where its respective senses would be. The conflation of word senses was perhaps less of a problem for Rooth et al., since the distributional evidence of predicate and argument was considered jointly.

Instead of using the distributional evidence to create an explicit classification, Grishman & Sterling (1993) used it to estimate confusion probabilities for words. These indicated the probability of one word occurring in the same contexts as another word, averaged over these contexts. With these confusion probabilities they then computed the smoothed probability for a novel combination. The smoothed probabilities were evaluated on a task of separating valid triples (<*predicate, relationship, argument*> from invalid ones. In these experiments, smoothing did appreciably improve coverage and recall. However, this was at the expense of the error rate. Using a manual classification appeared to achieve better results. However, further experiments with a larger corpus were suggested.

A related approach, also looking at common contexts, was the example-based or 'analogy-based' approach of Federici et al. (1997, 1999), Federici, Montemagni, & Pirrelli (2000). They

obtained co-occurrence data from a bilingual dictionary and used this to build up a data base of previous examples, the example base, where each example held the verbal predicate, syntactic relationship and argument head. Analogical families were drawn up between verbs which shared contexts. For example, verbs such as *fumare* (to smoke)[1] and *accendere* (to light) were linked by virtue of sharing a direct object, such as *sigarette* (cigarette). Such links between verbs then permitted inferences between novel predicate and argument combinations. For example, the co-occurrence of *accendere* and *pipa* (pipe) would be anticipated given a stored observation of the pattern *fumare-pipa/0bject*. Federici et al. used a weighted measure of the number of shared contexts to indicate the likelihood of novel combinations.

One criticism levelled at distributional approaches is that the output does not lend itself easily to symbolic interpretation, if that is required (Resnik, 1993a). For example, in a WSD task where predefined senses from a dictionary need to be assigned to text. Mapping from an automatic classification is difficult, particularly for incongruous classes. This is over and above the difficulties typically encountered when mapping between alternate symbolic taxonomies in the first place (Carroll & McCarthy, 2000). Smoothing methods are particularly problematic as a matrix of confusion probabilities is perhaps even further removed from symbolic interpretation than automatically produced classes are. Automatic methods are perhaps better suited when symbolic interpretation is not required. Indeed, for most NLP applications, symbolic interpretation is not necessary for the final output, although it is useful during system development. Results from the ROMANSEVAL evaluation (Federici et al., 2000) demonstrate that example bases can be readily combined with a predefined semantic classification: each example can have a hand-coded label attached. They divided their example base into a supervised and an unsupervised portion. The unsupervised portion contained in excess of 17,000 patterns for 3,858 verbs. The supervised portion contained an average of 6 labelled patterns for each of the verb senses of these verbs, giving an average of 32 labelled patterns for each of these verbs. Of course, this means we are again tied to human supervision which the other automatic classification systems aim to get away from.

Using Manmade Word Classes

The option of exploiting manmade taxonomies bypasses the computational expense and up-front effort required for automatic clustering. However, manmade resources bring with them other disadvantages. They rely on the introspections of lexicographers, even where the entries have been produced with recourse to citations from naturally occurring corpus data. They are therefore prone to human error and do not reflect distinctions within the sublanguage of a particular corpus.

The choice of manmade classifications available depends on the language in which they are required. The bulk of work acquiring selectional preferences has been done for English (with some exceptions, notably the analogy-based work in Italian, and application of the system described in Rooth et al. (1999) to a German corpus). In English, WordNet has featured strongly as a popular classification (Resnik, 1993a; Ribas, 1995a; Li & Abe, 1995; Abe & Li, 1996; Abney & Light, 1999). WordNet is an on-line thesaurus, organised by semantic relations rather than alphabetically. Words are classified by their part-of-speech, (noun, verb, adjective and adverb). They are then subdivided into small classes called 'synsets' where members are near synonyms of each other. These synsets are then linked together by semantic relationships such as hyponymy (nouns and

[1] The English translation for the Italian is provided in brackets.

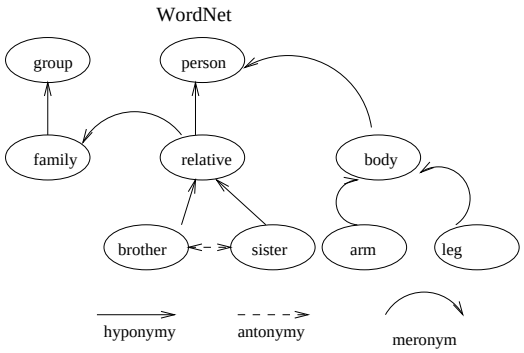

Figure 2.3: WordNet

verbs), meronymy (nouns), and entailment (verbs). Figure 2.3 illustrates some of the relationships between WordNet synsets.

Research that uses WordNet to capture semantic preferences has restricted itself to the noun hyponym hierarchy as nouns are the heads of the majority of slots for which preferences are sought. In the case of PPs, they are the heads of the noun phrase which is introduced by the prepositions. The hyponymy hierarchy lends itself most readily to interpretation for selectional preferences. The nouns are linked together by the hyponymy or IS-A relationship e.g. **brother** IS-A **relative** IS-A **person**. [2] When a verb shows a relationship with a superordinate class, for example, *believe* has a preference for **person** at the subject slot, then the relationship should hold for subordinate classes too, for example **brother**, **sister**, **preacher** etc... .

Other relationships have not as yet been used for selectional preference acquisition. Antonymy is potentially useful. However this relationship is usually covered by the hyponym relationship, since there is often a common superordinate parent within the hyponym hierarchy. It is less clear how the meronymy (PART-OF) relationship could be systematically used, since preferences at superordinate classes cannot necessarily be applied to subordinate classes and vice versa. For example, a *wheel* is PART-OF a *bus* however verbs such as *drive*, *park* and *reverse* which take *bus* as direct object, do not readily take *wheel*. Some verbs do take meronymy entailments, for example *touch* may take a direct object such as *arm* and also *person* of which *arm* is a subordinate class in the meronym hierarchy. To use the meronym hierarchy for automatic acquisition, one would need the system to determine the cases where the relationships were relevant.

WordNet has obvious appeal for use in acquisition of selectional preferences because of its widespread availability, with no licensing restrictions, and extensive coverage. Additionally it has the virtue of being organised by sense rather than form and when looking for semantic constraints it is the senses that are relevant. The electronic version of Longman's Dictionary of Contemporary English (LDOCE) (Procter, 1978) (version 1) makes use of a semantic classification which is used for the selectional restrictions on subject and object slots provided by lexicographers. This hierarchy is rather shallow in contrast to WordNet. The entire classification has 32 categories and some of these are combinations of a core set of 16. It may be advantageous to reduce unnecessary search

[2] Throughout this thesis, we shall refer to individual synsets in WordNet using one or two synonyms which are representative of the class. In WordNet, synsets are given unique numerical identifiers, since words can belong to more than one synset. These would not be meaningful to the reader. Where necessary, the relationship with other synsets will disambiguate which synset is being referred to.

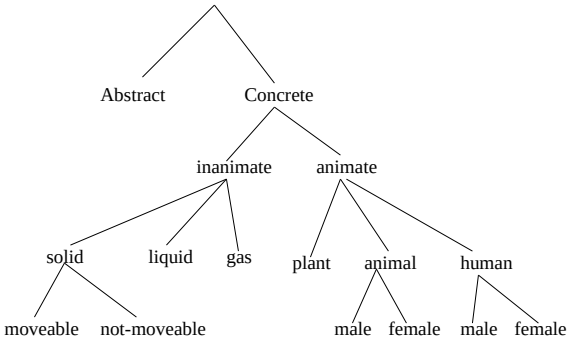

Figure 2.4: LDOCE semantic space

space by keeping to a simple hierarchy. However, it seems likely that a lot of specific predicates will not be adequately catered for. For example, given the 16 core categories depicted in figure 2.4 the direct object slot of *sail* would have to be accounted for by the **movable** class, when a more specific classification would be useful to distinguish, for example, *cars*, *stones* and *ships*.

There are now WordNet versions for some European languages other than English (Vossen, 1999). For other languages, producing a new man-made hierarchy is not an easy alternative. The coverage needed for even a restricted domain requires considerable human effort.

The noun hyponym hierarchy of WordNet is used as the representation medium for the preferences within this thesis. This makes our preferences prone to the human error inherent in the hierarchy and characteristic of any manmade resource. However, this is to some extent outweighed by the rigorous human effort that has gone into creating this useful taxonomy. WordNet has in excess of 60,000 classes in the hyponym hierarchy with over 88,000 word forms (version 1.5). Using current automatic classification methods for building a hierarchy of reasonable size would require considerable effort in post-editing to avoid incongruous classes and considerable processing time in the first place (Resnik, 1993a). The preferences we obtain are limited to the distinctions made within WordNet. Using corpus data does, to some extent, allow us to obtain preferences for the sublanguage of the corpus, since areas of WordNet that are not relevant to the domain have negligible frequency counts.

2.3 The WordNet Approaches

There is a common theme to the research acquiring selectional preferences using WordNet (Resnik, 1993a; Ribas, 1995a; Li & Abe, 1995; Abe & Li, 1996; Abney & Light, 1999). Preferences are sought for subjects, objects and prepositional phrases using the head noun of subjects and objects, and the head noun of the noun phrase for prepositional phrases. Indirect objects have been ignored presumably because they are rarer. They could be handled by the same mechanism used for subjects and direct objects. The prepositions in prepositional phrases are used alongside the verb as an anchor for the selectional preference.

The data in the corpus is used to populate the WordNet noun hyponym hierarchy with frequency counts. These counts are transformed into preference scores. Section 2.3.1 goes into more detail of how WordNet is populated with frequency counts. The hierarchy with preference scores

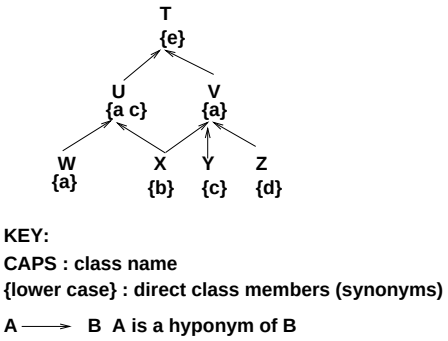

Figure 2.5: Taxonomy A

is used to represent the preferences as it stands (Abney & Light, 1999; Resnik, 1993a). Alternatively, the hierarchy is searched to find a set of disjoint classes (not ancestors of one another) to represent the preferences (Li & Abe, 1995; Ribas, 1995a). How this set is found is discussed in section 2.3.3. The method of generalisation depends on the preference scores used and these are described in section 2.3.2.

2.3.1 Populating WordNet with Frequency Information

Fundamental to all approaches using WordNet, is the need to produce a representation of the hierarchy populated with frequency or probability information. There are a number of different options for doing this. WordNet is organised by sense rather than word form and corpus data consists of the latter rather than the former. Moreover, in the hyponym hierarchy, synsets are linked to one another to indicate where one class shows an IS-A relation to another. When the word *chicken* is observed, how should the frequency distribution be divided between the various classes with direct membership and those with indirect membership? A probability distribution must sum to 1 but how should the probabilities at superordinate classes relate to the probabilities of their hyponyms?

To illustrate the difference in possible approaches, we will consider taxonomy A in figure 2.5. The UPPER CASE letters represent classes in the taxonomy. The lower case letters represent words with direct membership of the classes under which they appear. The arrows indicate the hyponymy relationship and any lower case member of a hyponym is an implicit indirect member of the superordinate classes. In this way the class **T** has *e* as a direct member and *a b c d* as indirect members. A class may have more than one member e.g. **U**, and an item may belong to more than one class e.g. *a*.

Resnik's approach doesn't distinguish between hyponymy and polysemy when estimating the frequency distribution. The frequency of a noun, in a given sample, contributes to all the classes the noun belongs to (classes(n)), regardless of direct or indirect membership. Furthermore, each frequency count is divided by the number of these classes to ensure that the sum of probabilities over the entire hierarchy equals one. Equation 2.2 defines his estimation of the frequency of a class (c).

Table 2.2: Resnik's frequency and probability distributions

CLASS	FREQ	PROB = $\frac{Freq}{5}$
T	$\frac{2}{4}+\frac{1}{4}+0+\frac{1}{3}+\frac{1}{1}=2.0833$	$0.41\dot{6}$
U	$\frac{2}{4}+\frac{1}{4}+0=0.75$	0.15
V	$\frac{2}{4}+\frac{1}{4}+0+\frac{1}{3}=1.08\dot{3}$	$0.21\dot{6}$
W	$\frac{2}{4}=0.5$	0.1
X	$\frac{1}{4}=0.25$	0.05
Y	0	0
Z	$\frac{1}{3}=0.\dot{3}$	$0.0\dot{6}$

$$\text{freq}(c) = \sum_{n \in \textit{nouns at or under}(c)} \frac{1}{|\text{classes}(n)|} \times \text{freq}(n) \tag{2.2}$$

In the work of Resnik and the other works described here, the estimation of class probabilities from the class frequencies is the straightforward maximum likelihood estimate:

$$\hat{p}(c) = \frac{\text{freq}(c)}{N}, \ \textit{where } N = \sum_{c' \in \textit{all classes}} \text{freq}(c') \tag{2.3}$$

A problem in Resnik's scheme arises from the lack of distinction between direct and indirect membership. The probabilities of hyponym classes are not propagated to their superordinates. [3] This gives rise to a number of anomalies because the contribution of a noun depends on the depth of the classes to which it belongs directly (direct classes(n)).

For example, given the taxonomy A in figure 2.5, if the string *a b d a e* is observed, the frequency and probability distributions would be as shown in table 2.2. The frequency contribution from each 'word' (lower case letter) is shown separately at each of the classes to which it belongs, directly or indirectly. So, for example, the frequency at class **V** is calculated according to its members which appear in the string. Its members are *a*,*b*,*c* and *d*. There are 2 *a*s in the string, and the frequency 2 is divided by 4, which is the total number of classes that *a* belongs to. There is one *b* in the string, and this is divided by the number of classes that *b* belongs to (4). There are no *c*s in the string so we have 0 from this member, and 1 *d*, which is a member of 3 classes.

The class probabilities in Resnik's scheme sum to 1 but there are anomalies that Resnik himself acknowledges. For example, **X** and **Z** both have members (*b* and *d* respectively) which occur the same number of times (once) but the classes end up having different frequency counts because of the difference in the number of superordinate classes above them. Resnik noted in his thesis that the assignment of class probabilities warranted further attention.

Ribas (1995a), in contrast, wanted to adhere to the maxim that the probabilities of all the possible senses of a noun (senses(n)) should sum to one, i.e.

[3] The superordinates in the hyponymy hierarchy are referred to as hypernyms in the WordNet literature (Fellbaum, 1998). This terminology has been criticised by Sampson (2000) because the Greek root for *name* is *onym*. The term for a superordinate class in the hyponym hierarchy should therefore be *hyperonym*. We will adopt Sampson's terminology.

Table 2.3: Ribas's frequency and probability distributions

CLASS	FREQ	PROB = $\frac{Freq}{5}$
T	$(2\times\frac{3}{3})+\frac{1}{1}+0+\frac{1}{1}+\frac{1}{1}=5$	1
U	$(2\times\frac{2}{3})+\frac{1}{1}+0=2.\dot{3}$	$0.4\dot{6}$
V	$(2\times\frac{1}{3})+\frac{1}{1}+0+\frac{1}{1}=2.\dot{6}$	$0.5\dot{3}$
W	$2\times\frac{1}{3}=0.\dot{6}$	$0.1\dot{3}$
X	$\frac{1}{1}=1$	0.2
Y	0	0
Z	$\frac{1}{1}=1$	0.2

$$\sum_{sense\in \text{senses}(n)} p(sense|n) = 1 \tag{2.4}$$

rather than the probabilities of all senses and hyperonym classes as is the case in Resnik's scheme. Ribas devised a local weighting scheme that would maintain this constraint. For this, he proposed a weighting where the frequency contribution of a noun to a class gets divided by the number of classes in which the noun has direct membership (direct classes(n)). Additionally, the probability of a hyperonym ($p(c)$) should be the sum of the probabilities of all hyponyms plus any probability it has by virtue of direct membership. In this way, the probability at the root of the hierarchy should equal one. He formalised this as a weight that is applied to the frequency count for each noun. The weight is specific to the noun (n) and class (c) and is the ratio between the total number of classes containing n beneath and including c, and the total number of classes directly containing n (direct classes(n)), this is shown in equation 2.5.

$$weight(n,c) = \frac{|\text{direct classes}(n) \in \text{classes at or under}(c)|}{|\text{direct classes}(n)|} \tag{2.5}$$

The unweighted frequency count at any class from any noun belonging at or beneath the class is simply the number of occurrences of that noun in the corpus. This is multiplied by the weight as in equation 2.5. The frequency of a class is then the sum of all the weighted frequencies of nouns in the corpus belonging by direct or indirect membership:

$$\text{freq}(c) = \sum_{n\in nouns\ at\ or\ under(c)} \text{freq}(n)\times\text{weight}(n,c) \tag{2.6}$$

The frequency and probability counts for the example string *a b d a e* are shown in table 2.3. The frequency estimate from each letter is again shown separately. The contribution of *a* is shown as the first component of the addition for the frequency estimation of each class it belongs to. It belongs to classes **W**, **U**, **V** and **T**. For each of these classes, the unweighted frequency 2 is multiplied by the appropriate weight. The denominator of the weight is 3 in all these cases, the number or polysemes of *a*. However, the numerator is 1 for **V** and **W** since only one direct sense falls under these classes. Meanwhile for **U** and **T** the numerators are 2 and 3 respectively reflecting the membership at and under these classes.

The basic scheme for frequency assignment used by Li and Abe[4] (Li & Abe, 1995; Abe & Li, 1996) is effectively the same as that of Ribas but just expressed a little differently. The frequency estimate for a class is initially calculated only with direct membership in mind, using the sum of the frequencies of each noun belonging at that class divided by the total number of classes that the noun directly belongs to. These frequencies are then cumulated up the hierarchy. The frequency estimation, and therefore probability estimation, is, in effect, the same as that of Ribas. However, the contributions of direct members are calculated by a separate process from those of indirect members and this is shown in equation 2.7.

$$freq(c) = \sum_{n \in nouns\ at\ class(c)} \frac{\text{freq}(n)}{|\text{direct classes}(n)|} + \sum_{n \in nouns\ under\ class(c)} \frac{\text{freq}(n)}{|\text{direct classes}(n)|} \quad (2.7)$$

With either formulation, the results are the same. As can be seen in table 2.3, the probability of the root T equals 1. Indeed, the sum of the probabilities across any set of disjoint classes covering all leaves, will be one.[5] Also, this method ensures that, for any noun, the sum of probabilities $p(c|n)$ at the direct classes equals one, whereas in Resnik's scheme it is the sum of probabilities from all classes for a noun (direct or hyperonyms) that equals one.

Abney & Light (1999) construed the task as one of generation. They placed WordNet within a hidden Markov model (HMM) and used the forward-backward algorithm, a specialised form of the (EM) algorithm (Dempster et al., 1977), to iteratively reestimate the probabilities of the transitions. The probabilities were propagated from the root to the leaves in the learning phase. There was a probability distribution over the hyponyms links of each node (class) that summed to 1. At every class, $p(child|parent)$ gave the probability of the transition to the child. If the child was a terminal, a noun lemma was emitted. The handling of ambiguity was not straightforward. The frequency count for an ambiguous word was split between its senses and the frequency counts were only considered with respect to the parent node of the respective sense. Unfortunately the model obtained when the basic algorithm converged still contained the ambiguity that was present in the data. To try and resolve ambiguity, Abney & Light modified the transition weights. Instead of using the unadulterated transition probabilities $p(t)$ as the transition weights for the next iteration, Abney & Light mixed in the uniform distribution $u(t)$ at each node. The uniform distribution was:-

$$u(t) = \frac{1}{number\ of\ children\ at\ each\ node} \quad (2.8)$$

They used a mixing parameter dependent on the total frequency count for the state. More specifically:

$$\varepsilon u(t) + (1-\varepsilon)p(t)), \quad where\ \ \varepsilon = \frac{1}{Node\ Frequency + 1} \quad (2.9)$$

For more frequent nodes, a larger portion of the empirical distribution was used in determining the weights. For less frequent nodes, a higher proportion of the uniform distribution was mixed in.

[4] We shall refer to the work in papers (Li & Abe, 1995) and (Abe & Li, 1996) as "Li & Abe" throughout, since the two pieces of work relate to each other and both involve the same two authors.

[5] This is with the exception of layers involving overlap of classes which are multiple-ancestors of the same descendent classes.

During the learning phase, the weight at less frequent nodes was not backed up by the evidence from the corpus and so weights in these areas were gradually diminished in subsequent iterations. Where there was no empirical evidence, the uniform distribution was applied. The basic idea was that senses in high frequency areas of the network would be preferred. This was introduced with the hope of disambiguating word senses, but it would have affected all nodes regardless of the polysemy of their descendants. The strategy penalised areas of low frequency. In addition to mixing in the uniform distribution, other modifications were also made to the basic forward-backward algorithm. These were devised to compensate for bias towards paths with more than one sense from the same lemma, and to compensate for the effect of path length and breadth (bias against long paths nodes with many subclasses). The authors acknowledged that the theoretical implications of altering the basic EM algorithm in this way have not yet been analysed. The modifications are perhaps an indication that HMMs are not well suited to representing the semantics of IS-A hierarchies. Nevertheless the approach is appealing because it seeks to model the stochastic process producing the training data.

In this thesis, we adopt Li & Abe's method of populating WordNet with frequency information. This accords with the semantics of an IS-A hierarchy where the root over all hyponyms has a probability of 1, covering the semantic subspace. The probability at classes is dependent on lemmas at that class and beneath, and the probability distribution over all classes of a noun, given that noun, sums to 1.

2.3.2 Measures of Preference

The WordNet approaches use the frequency distribution over the noun classes to obtain probability distributions, which are used for ranking analyses, either directly, or by using the probability or frequency distributions to obtain preference scores. The preference scores can also be compared to a threshold in tasks where a hard and fast decision is required.

Ribas has experimented with the largest range of preference measures (1995a), and we do not seek to repeat his work. We will, however, experiment with a selection of three measures suggested by others to further investigate their strengths and weaknesses.

The simplest measure is conditional probability $p(class|verb)$ (Abney & Light, 1999; Li & Abe, 1995, 1998; Li, 1998). Probabilities have the advantage that they can be readily combined in a sound way and packaged within a probabilistic system (Li, 1998). One disadvantage is that conditional probability does not take the quantity of the sample size into account. This means that low frequency data may not be catered for adequately. A further disadvantage in using conditional probabilities is that they do not control for the marginal $p(c)$ ('prior' or 'foreground' probability). For example, *insect* may have a low frequency as the subject of *fly*, but we need to contrast this with the prior distribution irrespective of the verb. When we see its low frequency in the corpus as a whole, its frequency as a subject of *fly* stands out. For this reason, many researchers have used scores based on mutual information, given in equation 2.10. [6]

$$MI(c,v) = \log \frac{p(c,v)}{p(c) \times p(v)} = \log \frac{p(c|v)}{p(c)} \tag{2.10}$$

[6] All logarithms are base two unless otherwise stated.

Mutual information is a measure often used within natural language processing which is intended to measure the association between two items. Typically these items are words but for measuring preference strength the item is usually a noun class (c) and verb (v). The measure contrasts the conditional probability of c given v with the prior (or marginal) distribution of the c irrespective of the verb.

A large prior probability radically affects the scores. A high preference score is not given where a noun class has a high probability of co-occurrence with a verb if the same noun class co-occurs with a similar probability in other contexts. Mutual information is typically higher for more specific classes since they are easier to match to a specific context. Resnik scaled mutual information by the conditional probability of the class given the verb. For each class, he calculated the selectional association as given in equation 2.11 (Resnik, 1993a):

$$A(v,c) = \frac{1}{\eta_v} p(c|v) \log \frac{p(c|v)}{p(c)} \tag{2.11}$$

$$\eta_v = \sum_c p(c|v) \log \frac{p(c|v)}{p(c)} \tag{2.12}$$

The divisor η_v provided a normalised measure (between 0 and 1) which took into account the strength of selection of the predicate across all classes. This was given by the relative entropy between $p(c|v)$ and $p(c)$ summed over all classes with respect to the target predicate. This normalising factor controlled for the selectional properties of the verb across all classes. Ribas adapted the unnormalised measure by experimenting with (i) the source of data for the prior distribution and (ii) the weighting he used to calculate the probability distributions. Abe & Li (1996) tried using a related 'association norm' measure which was in essence mutual information without the logarithm. They reported increased performance compared to both their previous implementation using conditional probability and to Resnik's selectional association measure on the task of structural disambiguation. In further work (Li & Abe, 1998) they reverted to using conditional probability. In his thesis, Li (1998) opted for probabilities because the theoretical foundations are clear and because of the ease of combining probabilities and manipulating them in a probabilistic system. Li also observed that simply altering the data used for the prior distribution can radically affect matters. [7] The structural disambiguation experiments reported in Li & Abe (1998) indicated that the association norm can be led astray by a poor estimate for $p(c)$.

As a terminological aside, we note that mutual information as used in natural language processing is not the same as the mutual information of information theory. This was pointed out by Dunning (1998). Since there is some relation, it is important to get the distinction correct. The mutual information that we have discussed so far, which Dunning called 'single celled mutual information', is actually a single data point in the mutual information of information theorists ('average mutual information'). The latter measures the association between both variables over all values as indicated in equation 2.13:

$$\mathrm{MI}(X,Y) = \sum p(x,y) \log \frac{p(x,y)}{p(x)p(y)} \tag{2.13}$$

[7] Personal communication.

We will continue to use mutual information for the single celled type since this is the term used within NLP.

Dunning's point was more than one of terminology. His concern was that the use of mutual information by computational linguists is generally misguided since the measure is poor at handling low frequency items. If a noun occurs next to a particular verb and has a low frequency in our corpus we can be fooled by a high association score simply because we do not have a large enough corpus to prove that it is genuinely rare. Dunning recommended using log-likelihood ratio (LLR) tests instead. These are referred to in some texts as G^2. Dunning demonstrated the utility of the binomial version of this test for finding highly associated bigrams (Dunning, 1993). The advantage is that such tests they take the sample size into account and can better detect relationships between word pairs which are unusual when considering the prior distribution. LLR selects pairs which have unexpected relationships rather than trying to measure the degree of association. In this way, it acts as a filter to detect relationships that would be unlikely to have occurred by chance rather than simply those that have a high association score. For example, a pair of words that occur together once may have a strong association score, but the result does not indicate how likely this is to have occurred by chance. One of the words in the pair may have only occurred this one time in the entire sample and its co-occurrence with the other word might be quite coincidental.

Many of the tests used in computational linguistics for identifying relationships between words are inappropriate because they mishandle low frequency items. As is well known, low frequency items are common-place in natural texts, in accordance with Zipf's law (Zipf, 1935). Many statistical tests, such as z-score tests, assume that the underlying variables are normally distributed. Dunning recommended LLR because it does not depend so highly on assumptions of normality and therefore allows the comparison of rare events with common ones. It is a parametric test, but it assumes the distribution of the generalised log-likelihood ratio and Dunning reports that it can be applied to much smaller texts than tests which are based on the normal distribution. It should be noted that, whilst mutual information along with other metrics are criticised for over-estimating low frequency items, the binomial LLR has been observed to go the other way and underestimate these (Dunning, 1993; Ribas, 1995a; Pedersen, 1996).

Dunning demonstrated that LLR [8] is useful for finding significant co-occurrences between words. This measure can also be used for finding selectional preferences between predicates and arguments. The task can be construed as one comparing two binomial distributions. For example, assume we are considering head nouns, instead of noun classes, appearing in the direct object slot of the target verb. The counts collected are:

1. verb and and noun together (k_1)
2. verb with any other noun ($n_1 - k_1$)
3. noun with any other verb (k_2)
4. other noun with any other verb ($n_2 - k_2$)

[8] Hereafter, references to LLR refer to the binomial version of this test.

The LLR statistic is :

$$\begin{aligned} -2log\lambda = 2[\quad & logL(p_1,k_1,n_1) \\ & +logL(p_2,k_2,n_2) \\ & -logL(p,k_1,n_1) \\ & -logL(p,k_2,n_2)\,] \end{aligned} \tag{2.14}$$

where

$$logL(p,n,k) = k \times \log p + (n-k) \times \log(1-p)$$

and

$$\hat{p}_1 = \frac{k_1}{n_1},\ \hat{p}_2 = \frac{k_2}{n_2},\ \hat{p} = \frac{k_1+k_2}{n_1+n_2}$$

The statistic tests the likelihood that the probabilities p_1 and p_2 are the same, i.e. whether the probability of the noun occurring given that the verb has occurred is the same as the probability that the noun has occurred given that the verb had not occurred. If this were the case, then LLR would be 0. A large LLR indicates that any association is unlikely to be due to chance. LLR is not a measure of the strength of the relationship and it can indicate a relationship in either direction, i.e. $p_1 > p_2$ or $p_1 < p_2$. The former would be the case for a pair of words which occur together more than expected, whereas the latter indicates that the pair co-occur less than anticipated.

The examples below help to illustrate the differences between the association norm measure (Abe & Li, 1996) and LLR. Suppose that *hedgehog* is seen once as direct object to *eat* and not with any other verb. In contrast *ketchup* is seen twice with *eat* and twice with other verbs. The association score for *ketchup* halves in value, meanwhile LLR increases with the additional evidence. *Sandwich* is seen four times with *eat* and four times with other verbs. The association scores stays the same as for *ketchup* since the ratio is the same but LLR is more than doubled thanks to the extra evidence.

eat hedgehog	Ass =	$\frac{1}{100}$ / $\frac{1}{100000}$	= 1000
	LLR =	$\frac{1}{99+1}$ vs $\frac{0}{99900+0}$	= 19.9
eat ketchup	Ass =	$\frac{2}{100}$ / $\frac{4}{100000}$	= 500
	LLR =	$\frac{2}{98+2}$ vs $\frac{2}{99898+2}$	= 31.9
eat sandwich	Ass =	$\frac{4}{100}$/ $\frac{8}{100000}$	= 500
	LLR =	$\frac{4}{96+4}$ vs $\frac{4}{99896+4}$	= 64.0

Table 2.4: Contingency table for *eat sandwich*

	sandwich	¬ sandwich	totals
eat	4	96	100
¬ eat	4	99896	99900
totals	8	99992	100000

LLR tests provide a viable alternative to the mutual information based association measures but there are others. Pedersen recommended the use of Fisher's exact test (1996) for finding dependent bigrams. Both tests are calculated with reference to contingency tables drawn up for the data observed. For the binomial (bigram) case, there are two rows and two columns representing the two variables as illustrated by the example in table 2.4. The calculation of LLR is given above.

Fisher's exact test is performed by exhaustively computing the probability of every contingency table that would lead to the marginal totals observed in the data. This can be extremely costly for some tables. The primary advantage this test has is that it does not assume an underlying distribution. It is therefore more reliable in cases where we cannot be sure what the underlying distribution of the data sample is like. In Pedersen's experiment bigrams were ranked according to (i) Fisher's exact test (ii) χ^2 approximation to LLR (iii) χ^2 approximation to Pearson's X^2 and (iv) the t-test. For the dependent bigrams, the ones where the words were clearly related to one another, the significance values were nearly identical. For the independent bigrams, there were differences. LLR tended to be more conservative and overstate independence. There were also examples of moderately independent bigrams where LLR went the other way and indicated a lower likelihood of independence than Fisher's exact test. Interestingly, the ranking for Fisher's exact test and χ^2 approximation to LLR came out the same in this experiment. Although Fisher's exact test is probably more reliable than LLR, since it does not assume any underlying distribution, we are not convinced that the slight differences in performance demonstrated by Pedersen warrant the lengthy computations. There is a further advantage in using LLR since it relates to the method we use for finding the right level of generalisation. We shall expand more on this in section 2.4.2.

Although LLR is a better measure for finding relationships between words than measures based on mutual information, using a class-based approach should reduce the effect of low frequency data. Preferences are typically acquired at a level of the hierarchy where the classes cover many words. However, it is unclear whether we would benefit from a better motivated measure in a class-based approach since many classes will have low frequency. Thresholding is not a tidy way of handling low frequency data. A method better equipped to deal with these items is desirable.

In a class-based approach, the nouns are replaced by noun classes in the calculations. Ribas has experimented obtaining preferences using a variety of measures including LLR. LLR indicates a relationship in either direction (preference and dispreference). In his use of LLR, Ribas searched for classes with a high LLR score only where the score reflected a preference. In his experiments, LLR achieved similar results to the association measure, although it did seem to be more conservative and more accurate.

All scores have advantages and disadvantages. Conditional probability permits integration in a

probabilistic system. Association measures give us an intuitive measure of association, taking into account the frequencies of the classes regardless of context. LLR is better equipped for handling low frequency data. In this thesis, we experiment with all these scores. This allows some choice for diathesis alternation identification techniques and provides a contribution to research into the relative merits of these scores for preference acquisition. We use a signed version of LLR indicating the polarity of the relationship. Unlike Ribas, we seek classes with high absolute values, since high negative scores should be as informative as high positive scores.

2.3.3 Preference Output

Having settled for an approach based on WordNet, a method of populating the WordNet hierarchy with frequency information, and some preference measures, we now need a method of extracting the classes and scores to represent our preferences. But which classes should be included? Resnik's early work searched for the best class for each verb (Resnik, 1992). However, if a best first search is conducted this will lead to getting stuck on local optima (Ribas, 1995a). In addition, one needs to allow for cases where the verb has preferences in more than one area, which is especially common where the verb form is polysemous. In later work, Resnik (1993a, 1997) retained the verb specific scores for the entire hierarchy for WSD and structural disambiguation applications. Abney & Light (1999) also kept the full hierarchy, in the guise of a HMM for each verb. The storage overhead of this must be balanced by the requirements of the application. For speech recognition, one might want estimates for $p(lemma|class)$ and $p(class|verb)$ and the expense may well be warranted. For other applications, such as ranking word senses or PP-attachments, the expense may not be justified.

Other researchers have sought a set of disjoint classes that cover the hierarchy. Ribas retained the highest scoring disjoint set of classes using his adaptations of Resnik's selectional association measure to find these classes.

Li & Abe (1995, 1996) used a representation of selectional preference bearing a resemblance to that of Ribas. This comprised a disjoint set of classes across WordNet with attached scores. Li & Abe devised a novel method of finding the best generalisation level, a method that has some clear theoretical underpinnings. We adopt Li & Abe's method of acquiring preferences, with some modifications.

Generalisation with the Minimum Description Length Principle

Rather than modify the association measure, as Resnik and Ribas did to get the correct level of generalisation, Li & Abe (1995, 1996, 1998) used a principle of data compression from information theory to find the appropriate level of generalisation. This principle is known as the Minimum Description Length (MDL) principle. In their approach, selectional preferences were represented as a set of classes or a 'tree cut' across the hyponym hierarchy which dominated all the leaf nodes, representing the noun senses, exhaustively and disjointly. Thus, a tree cut was a set of classes across WordNet that together covered all the leaves, and where none of the classes were ancestors of any other class in this set. The set of root classes of WordNet was just one possible tree cut. To ensure all noun senses occurred at or under a class on any tree cut, Li & Abe only used a shallow version of WordNet by prunning at classes where a class member had occurred in the corpus data. We comment on this further in section 2.4.1. The tree cut featured in a model, termed a 'tree cut

model' (TCM) which identified a score for each of the classes in the cut. This score was obtained from the corpus data and was used to indicate the preference for the class, and noun senses falling under that class.

The MDL principle (Rissanen, 1978) bears a resemblance to Occam's Razor. This is attributed to William of Occam who is credited with saying:

> "Entia non sunt multiplicanda praeter necessitatem"

The rule states that entities should not be multiplied needlessly. This is interpreted to mean "if two theories explain the facts equally well then the simpler theory is preferred" and is known in logic as the law of parsimony. The MDL principle reiterates this rule in information theoretic terms by stating that the best model is the one which has the shortest description length when measured in bits. The description length has two components:

1. The model description length - the number of bits to describe the model
2. The data description length - the number of bits to encode the data in the model

The best model is the one which minimises the sum of these two components. This provides a compromise between a clear and simple model and one which matches the data well.

Li & Abe (1995) initially devised a method for calculating a description length for a probabilistic TCM using the conditional probability $p(c|v)$. We will hereafter term this a probabilistic tree cut model (PTCM). Along with the calculations required for the description length, they provided an efficient algorithm for searching WordNet to find the cut model with the minimum description length. The actual encoding (in bits) was not actually produced since the optimal TCM was all that was required. Later, in (1996), they went on to devise a method for producing association tree cut models (ATCMs) where the association norm measure $\frac{p(c|v)}{p(c)}$ was used to indicate preference strength, and was also used for calculating the description length.

For the PTCMs the model description length was the number of bits to describe the cut under consideration, plus the number of bits to encode the parameters. Li & Abe opted for a uniform encoding of the cuts so that the number of bits to describe each cut was the same. This corresponded to assuming all cuts are equally likely a priori. The number of bits to describe the parameters was calculated by:

$$\frac{k}{2} \times \log |S| \tag{2.15}$$

where k was the number of free parameters (the number of classes in the cut minus one) and $|S|$ was the sample size. This was known to be the most efficient way of describing probability parameters (Rissanen, 1986).

Li & Abe's PTCMs (1995, 1998) provided a probability distribution across all leaves (noun senses). The data description length assumed that the data was encoded using the probabilities:

$$\text{number of bits} = -\sum_{n \in S} \log \hat{p}(n) \tag{2.16}$$

The probability estimates ($\hat{p}$) for a noun (n) were calculated by dividing the probability estimate for each class on the cut that the noun belonged to, directly or by virtue of its membership

of a hyponym class, by the total number of nouns at or under that class. We have described Li & Abe's method of estimating the probability of a WordNet noun class above in section 2.7 on page 25.

Thus, the description length for the PTCMs is defined as:

$$PTCM\ DL = \frac{k}{2} \times \log|S| - \sum_{n \in S} \log \hat{p}(n) \tag{2.17}$$

Rather than searching for the classes with the highest score, MDL was used to search for the classes which made the best compromise between explaining the data well by having a high probability, contributing towards a low data description length, and providing as simple (general) a model as possible, thus minimising the model description length.

In further work (1996), Li & Abe took the marginal distribution ($p(class)$) into account, as many other researchers do when obtaining preferences. They switched to ATCMs which used the association norm. The association norm is not itself a probability, but is obtained using probabilities. This complicated the calculations of the description length for the ATCMs. Li & Abe used the method of calculating description lengths for probability distributions and then calculated the relative cost of the ATCMs by using the identity:

$$p(n|v) = A(n,v) \times p(n) \tag{2.18}$$

A TCM was first obtained for the prior distribution ($p(class)$), using the description lengths for the TCM from the sample of head noun instances for the target slot irrespective of verb. Then, a verb specific ATCM was derived using this fixed prior model and the verb specific data. This was done as though the ATCM was a by-product of estimating a pair of TCMs that represented the conditional data.

The model description length for the ATCM was calculated as for the TCM, using the sample size for the data specific to the target verb. The model description length for the TCM (prior distribution) was not incorporated into this since it was fixed by the time the ATCM was determined.

The data description length can be written in two parts:

$$\sum_{n \in S} -\log p(n|v) = \sum_{n \in S} -\log A(n,v) + \sum_{n \in S} -\log \hat{p}(n) \tag{2.19}$$

the sample (S) was the data for the target verb, whereas the model for the prior distribution was estimated using the entire data set for the specified slot, irrespective of the verb. The summations were performed over all nouns (n) in the sample. These were represented by the classes in the TCM being examined. The second term was dropped as the TCM for the conditional distribution was not actually required and this term did not affect selection of the optimum ATCM. To obtain the first term, Abe & Li outlined a method of estimating the probability at a particular class given the fixed prior TCM either above or below it. If the class fell under a class on the cut then a portion of probability estimate on the cut was assigned to this class. This proportion was determined by the number of noun senses at or under this class, divided by the number of noun senses under the class on the cut. If the class fell above a set of classes on the cut, then the probability estimate was taken as the sum of the estimates from these classes. Where the class fell on the cut the estimate was taken direct from the cut model.

Thus, the description length for the ATCMs is defined as:

$$ATCM\ DL = \frac{k}{2} \times \log|S| - \sum_{n \in S} \log A(n,v) \tag{2.20}$$

Li & Abe provided an algorithm which searched the WordNet hyponym hierarchy efficiently by comparing cuts locally within subtrees. The search proceeded from the leaves comparing the cost of a cut at the leaves to a cut at the parent of these leaves. The best cut was propagated upwards. At each class (the root of a subtree), the optimal cuts at or beneath each of the children were appended together. The combined cost of these appended cuts was compared to the cost of a cut at this class (subtree root) and the optimal cut was propagated up. The process continued until the final comparison was made between the root of WordNet [9] and the best cut found beneath this. The algorithm guaranteed finding the model with the minimum description length.

Li & Abe's method using MDL to find the correct level of generalisation appeals to us because its clear theoretical underpinnings guarantee that one will find an optimal model in terms of the shortest description length. A method of generalisation is desirable given the large quantities of data involved. We use WordNet to define our semantic space because of its availability and coverage. Once extensive classifications can be built automatically then it would be interesting to compare these, particularly in terms of their ability to exploit the semantic space of a sublanguage. Li & Abe contrasted the use of automatic and manmade classifications on a PP attachment disambiguation task (1996). They demonstrated that WordNet achieved a greater level of coverage than their automatically clustered taxonomy, but that their automatic classification produced better precision. Optimal results were obtained by harnessing the two together.

There are some outstanding issues within selectional preference acquisition and we pick up on a few of the pertinent ones, some specific to Li & Abe's approach to acquisition, in the following section.

2.4 Modifications to the Basic Approach

2.4.1 Alterations to WordNet Structure

Li & Abe's approach was motivated by a theoretical standpoint on how best to obtain the correct level of generalisation. However, they required some changes to WordNet in order to make this work. Their scheme required that all noun senses fell under any potential cut. This ensured that all items are covered by the probability distribution on the cut. For this reason, all noun senses had to be placed at the leaves of the hierarchy. In order to meet this requirement, modifications had to be made to the original structure of WordNet where nouns are found at all levels in the hierarchy. To adhere to this constraint, Li & Abe pruned the hierarchy at classes where a direct class member (one of the nouns in the synset) had occurred in the data. However, this strategy can give rise to overly-general preferences when the data contains an argument which occurs higher in the hierarchy than the prototypical arguments of the verb. For example, in the data for lexicon A collected from the BNC, we observed the tuple $< build, direct\ object, entity >$ (<verb,slot,lemma>). **Entity** is one of the 11 root classes of the noun hyponym hierarchy of WordNet. When the hierarchy

[9] We have introduced a dummy root as a parent of the eleven root classes in the WordNet hyponym hierarchy. We refer to this hereafter as the dummy root.

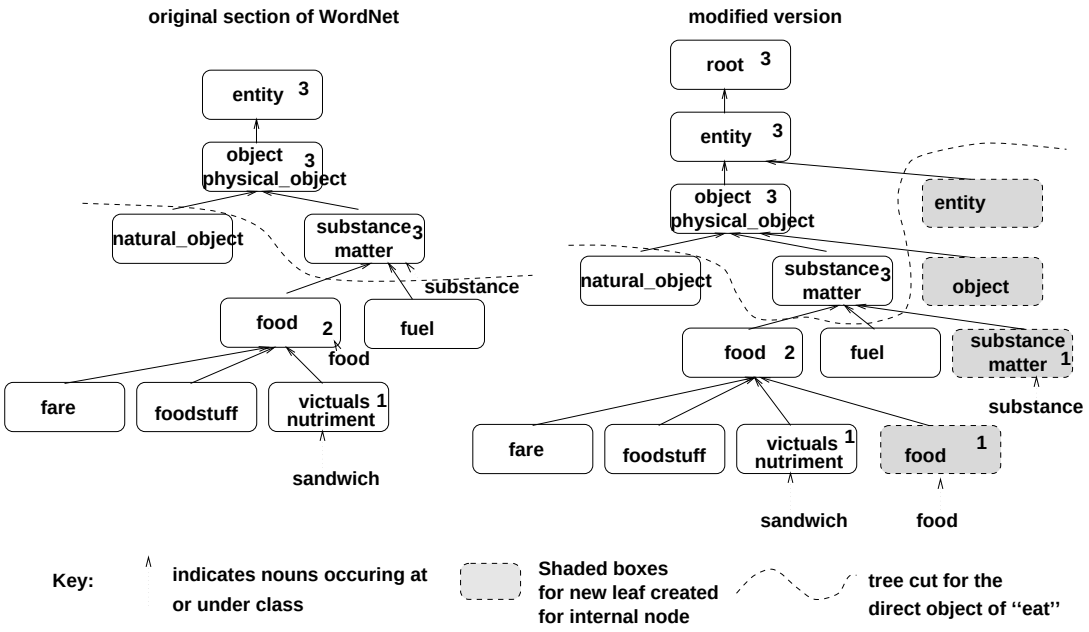

Figure 2.6: Creating leaves for internal nodes

is pruned at such a level, much detail is lost on the types of entities that are built. For example, no distinction can then be made between the **object** class and the **life form** class.

Li & Abe's strategy of pruning the hierarchy at all direct occurrences in the data entailed an extremely shallow version of the hierarchy and overly-general preferences. In the work described in this thesis, as an alternative, new leaf classes were created for every internal class in the WordNet hierarchy. This means that terminals only occur at leaves, but that the detail of WordNet is left intact. The frequency count attributed to nouns listed at the internal node is placed instead at the new leaf and then cumulated up the hierarchy in the usual way.

To illustrate, look at the transformation of a small portion of the hierarchy in figure 2.6. The class frequency distribution for three heads, *substance*, *sandwich* and *food*, from the direct object slot of *eat* is shown by the numbers in the right-hand corner of the class boxes. [10] All classes without numbers have zero frequency with respect to this small sample. In the unmodified version of WordNet, it is quite possible to have a cut which does not cover all the noun senses and so the probabilities of the classes along the cut would not sum to one. This is because nouns such as *substance* occur at internal classes which can occur above a candidate cut. In our scheme the frequency, and therefore probability, contributions from such nouns are moved down to newly created leaves and so the probability axiom ($\sum_{c \in classes\ on\ cut} p(c) = 1$) is maintained. In contrast, the strategy adopted by Li & Abe would be to prune the tree at the **substance** class which would result in a shallow tree with no possibility of distinguishing between a preference for **food** or **fuel** at the direct object slot of *eat*.

[10] For the purposes of this example assume that these words are monosemous.

Thresholding

Alongside the issue of keeping terminal word senses at leaves, is the issue of thresholding. Li & Abe style pruning also ensured that low frequency areas of WordNet were not investigated. Li & Abe additionally used thresholding to discard classes with a probability estimate less than a threshold (0.05 for the PTCMs) (1995, 1998). In the ATCMs, thresholding (at 0.01) was only performed on the model for the prior distribution, the reason for this was not given. Presumably the ATCM cuts were typically more general than the prior model and therefore more compact. Crucially, Li & Abe performed this thresholding after the description length calculation, so that all classes were considered in the costing. Li & Abe stated that thresholding was performed for clarity and to remove some of the noise resulting from factors such as erroneous word senses. The main effect was to remove low frequency classes from the final cut.

We attempted to continue the use of Abe & Li (1996) style thresholding *alongside* our strategy of creating new leaves. However, this gave rise to unacceptable performance problems: the TCM for the prior distribution had not finished after 12 hours on a Sun Ultra. Instead, for the TCM used for the ATCM a new method of thresholding was tried in which the subtrees of classes with probability less than the threshold were not explored, and not included in the cost of the cut. This altered the structure of WordNet and reduced the search space by removing low frequency items.

A significant disadvantage of performing thresholding, with either method, is that low frequency areas are not covered by the cut. For the probabilistic models using the conditional data, we do not adopt thresholding. Thresholding could be applied as in (Li & Abe, 1995, 1998) by simply removing the classes when using or describing the cut. Since probabilistic models do not require a prior model, they are simpler and quicker to produce. Avoiding thresholding provides a further advantage since all areas of WordNet are covered, although this leaves the models prone to the noise from low frequency items.

Experiment

We compared ATCMs obtained (i) with Li & Abe's method of pruning and thresholding and (ii) by adding our new leaves for internal nodes and our system of thresholding. The ATCMs were obtained from lexicon A (a SCF lexicon produced from 10.8 million words of parsed text from the BNC). The verbs were all those which were observed in a random sample of 500 sentences from the Susanne corpus (Sampson, 1995). The preferences were acquired for the direct object slot.

We leave formal evaluation for chapter 4. On informal comparison, we observed that both the Li & Abe models for the prior distribution and the final ATCMs were substantially more general. Using Li & Abe's method of pruning and thresholding on this BNC data, we obtained a model for the prior distribution with 11 classes. This contrasted with 30 classes for the prior model using our internal nodes and thresholding. A small portion of the two cuts in the vicinity of the **entity** class is shown in figure 2.7. The detail in the prior model was matched by considerable specialisation when the ATCM models were produced.

For example, the ATCM for the direct object slot of *build* using our approach included the hyponym class **object** with a high association score (4.4) contrasted with a low score of 0.05 for **person_1**.[11] Meanwhile, in our reimplementation of Li & Abe's ATCM using the same data, we ob-

[11] Newly created leaves are indicated in the text with one of the words that were stored at the internal class plus a ⊥ suffix.

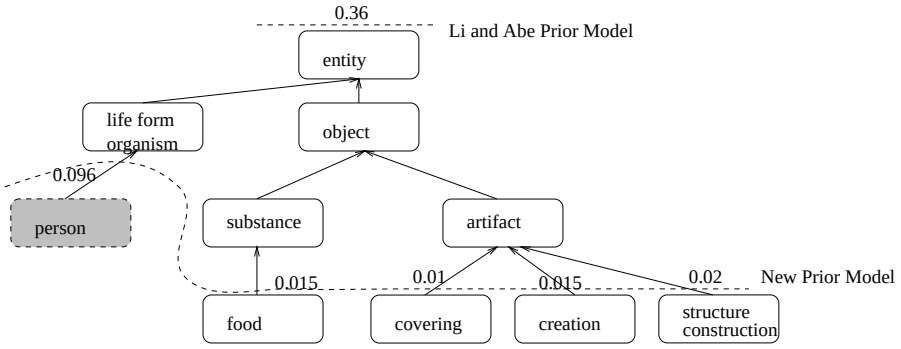

Figure 2.7: Tree cut models for the prior distribution.

tained a cut at the **entity** node with an association score of 1.2. This did not permit discrimination of nouns which occur at hyponyms of this class.

For our experiments, we placed a threshold on the number of argument head instances that were required before preference acquisition could be attempted for a given predicate and slot. The threshold applied to the number of these instances that were classifiable in WordNet and was set at 10. There were 395 verbs, within the sample of 500 sentences, that met this requirement for the direct objects slot. Some of these verbs were cut at the dummy root indicating that the MDL method did not find a preference model below this was warranted given the data. We will hereafter refer to a cut model at the dummy root as a 'root cut'. This occurred to a greater extent for the Li & Abe style pruning than with the method of adding leaves for internal nodes. For Li & Abe style prunning there were 209 root cuts (more than half the sample). Using our method, the number of root cuts was reduced to 101. It could be that a more conservative approach, which does not produce discriminatory preference models unless there is strong evidence, is beneficial. However, when we looked informally at the data we concluded that many of these root cuts were unnecessarily uninformative. For example, using Li & Abe pruning and thresholding, the model for the direct object slot of the verb *melt* was cut at the root. This verb clearly does exert some preference for the semantic type of direct objects that it occurs with. Our method provided us with an ATCM with a high association score (36) at **substance**.

For PTCMs we also used our strategy of creating new leaves for internal nodes. This again allowed more detailed cut models than those created using Li & Abe style prunning. Even using this strategy on informal evaluation, the cuts for the PTCMs appeared to be more general than our ATCMs. We will illustrate these differences at the end of section 2.4.2. We did not apply thresholding to the PTCMs since they involve simpler computation than ATCMs and can feasibly be produced without this. It would be straightforward to apply thresholding in the Li & Abe style at a later stage, if required. Removing low frequency data feels a little like sweeping the dirt under the carpet. Instead, we looked into incorporating the LLR statistic, which has been reported to be well equipped for handling low frequency data.

2.4.2 LLR Models

The LLR statistic described in section 2.3.2 provides a viable alternative to the association scores. It is better equipped to deal with low frequency items, whilst being computationally feasible. We

used scores to obtain LLR tree cut models (LLRTCMs) with signed LLR scores along the classes on the cut. The score associated with each class on the cut was given a sign (positive or negative) to indicate the direction of the preference, i.e. whether the class on the cut had occurred with the verb in the specified slot more or less than expected. These LLR scores attached to the classes on the cut represented the preference score for the class, and all descendant leaves under the class. Since LLR is not a probability and its calculation is clearly more complicated than the association norm measure, it was not straightforward to arrive at a description length for these models. LLR tests do bear a resemblance to full MDL methods (Dunning, 1998), and it is this resemblance that we exploited when calculating the relative costs of our models.

LLR tests can be thought of as a specialisation of the MDL principle (Dunning, 1998). They can be used to approximate the difference (measured in bits) between the encoded size of the observed data to the encoded null hypothesis. This is the special case where the two models we are deciding between are the null hypothesis (no difference between p_1 and p_2) and the unrestricted (alternative) hypothesis (that there is an inequality in either direction). In contrast, our task involved choosing from a variety of TCMs the one that made the best compromise between the cost of the model and the cost of encoding the data in the model. In our algorithm for finding the TCM, we compared the cut at each node, with the best found in the subtree beneath. These two alternative TCMs represent two possible models to represent the unrestricted case. To obtain a preference model, we wanted to maximise the significance of the difference between our model and the null hypothesis. When comparing two cuts, the one with the larger LLR was preferred.

LLR can observe relationships in either direction without distinguishing them. However, a sign can be used to indicate the direction of the relationship between the verb and argument. This allowed LLRTCMs to express dispreference as well as preference. The association norm measure, $\frac{p(c|v)}{p(c)}$, indicates negative associations with a score between 0 and 1. However, these are unreliable because of the poor handling of low frequency items discussed in section 2.3.2. Although we included a sign on the LLR scores in our LLRTCMs, when incorporating dispreference into our calculations, it was important that the cut model was not penalised at levels where the preferences and dispreferences cancelled each other out. To avoid this, the costing was performed with no indication of the direction of the relationship. The model with the largest value of LLR was favoured. When minimising the description length, we placed a minus sign in front of the LLR value (equation 2.21), since we wanted the cut with the largest absolute LLR value. The scores that we attached to the classes on the LLRTCM included a sign to indicate the direction of the relationship.

LLR can be used as a heuristic stand in for full MDL methods (Dunning, 1998). From preliminary experiments, we found that choosing a model on the basis of this measure alone resulted in very detailed cuts. This drastically affected the time taken to acquire the LLRTCMs and also resulted in a poor level of generalisation. The previous model description length was added to make a better compromise between the cost (no longer a true description length) of the data, and the cost of the model. We calculated the cost of a cut as:

$$\text{LLRTCM description length} = -\text{abs}(\text{LLR}) + (\frac{k}{2} \times \log|S|) \tag{2.21}$$

Our method of calculating the relative costs of the LLRTCMs no longer conforms to MDL, since the cost does not reflect the actual cost of an encoding. It does, however, bear some resemblance in

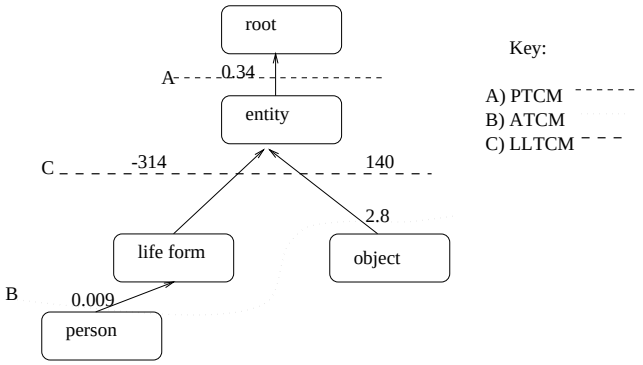

Figure 2.8: Cut models for *produce* direct object slot

Table 2.5: Number of root cuts, for different models

model	# root cuts
PTCM	141
ATCM	101
LLRTCM	63

that the relative costs of the alternative models are considered. We investigate whether the benefits gained by choosing a measure better equipped to handle low frequency items compensate for the departure from a clear calculation of description length.

Informal evaluation of the difference between the cuts obtained for the LLRTCMs, ATCMs and the PTCMs reveal some differences. These informal evaluations are made on the models created for the object slot of the 395 verbs with data from the lexicon described in section 2.4.1. Neither the PTCMs nor the LLRTCMs employed thresholding. Thresholding was only employed when obtaining the TCM for the prior distribution and only the ATCM required this. Thresholding is likely to allow more specific cuts since lower probability detail further down the tree is ignored and therefore a deeper cut does not incur such a high penalty. The PTCMs do tend to have less specific cuts than the LLRTCMs and ATCMs. This is exemplified in figure 2.8 which shows a small portion of the cut for the direct object of *produce* in the vicinity of the **entity** class. The PTCM cuts at the **entity** node itself which is not very informative. Table 2.5 compares the number of cuts at the root of the hyponym hierarchy for the sample of 395 verbs with an object slot in our sample.

It is less clear cut whether LLRTCMs are more specific than ATCMs. They have less root cuts, but for many examples the cuts are less specific (higher in the hierarchy). From informal evaluation, it appears that they more frequently display preferences but are more conservative and do not go into a lot of detail unless there is a lot of evidence. The example in figure 2.8 shows a LLRTCM cut falling between the PTCM and the ATCM. The LLRTCM is able to show a strong dispreference for the **life form** class at the direct object of *produce* which we believe is a strong asset of the model. ATCMs show negative associations less reliably because of their poor handling of low frequency items.

The preferences from the LLRTCM at the object slot of *feel*, as illustrated in figure 2.9, are much

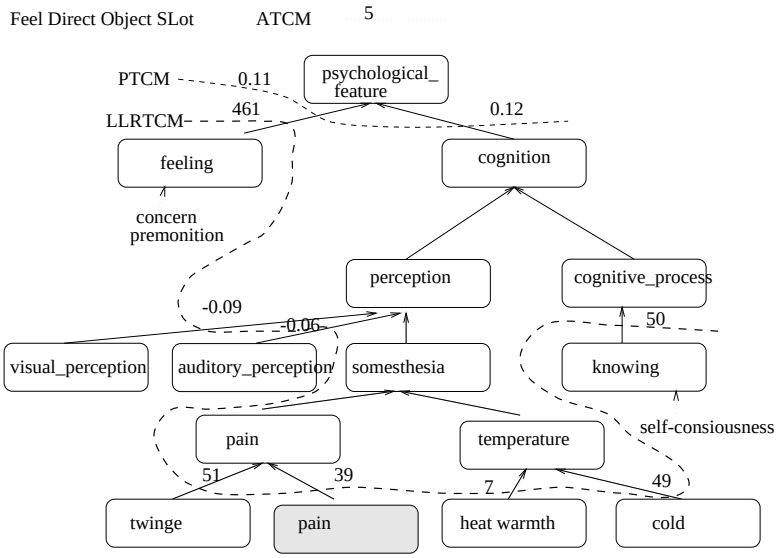

Figure 2.9: Cut models for *feel* direct object slot

more specific than the ATCM preferences. The strength of the preferences in the **psychological feature** are shown by the detail and strength of the cut in this area. The ATCM has a very general cut at **psychological feature**, one of the roots in WordNet. This does not allow any distinction for nodes beneath this. For example, no distinction can be made between the subclasses concerned with sound, smell, sight and touch. The LLRTCM, on the other hand, has a preference for the appropriate classes concerned with sensations of touch but not the other sensations. The PTCM on this occasion is more specific than the ATCM, although less specific than the LLRTCM.

Although a positive score on the LLRTCM tends to coincide with an association score above 1, this is not always the case. For example, the verb *produce* has the class **location** on both ATCM and LLRTCM cut models. In the latter, the score given is -23 indicating a dispreference whereas the ATCM shows an association score of indicating a preference 1.4. The association score is inflated because of a low prior probability at the **location** class. The PTCM records a low conditional probability of 0.01.

2.4.3 Word Sense Disambiguation

Acquisition of selectional preferences has typically been carried out on noun lemmas extracted from the WSJ parsed as part of the Penn Treebank II. WSD has not generally been performed on the input data, although Ribas (1995a) has attempted to investigate the benefits by comparing acquisition from the small sample of hand tagged text found in the SemCor data (Miller et al., 1993a). In the work of Li & Abe, Resnik and Ribas, frequency credits have been divided between all the senses of the words in the hope that the noise from erroneous senses is diminished with sufficient data. Erroneous senses were, however, reported as a common source of error, particularly that of over-generalisation (Ribas, 1995a). Li & Abe hoped to alleviate this problem by placing a threshold on the class probabilities. It is likely that acquisition from sense tagged data would provide more accurate frequency estimation, and that this will in turn produce better results on application of the preferences to NLP tasks.

WSD is a vast area of research in itself. Selectional preferences have long been linked with sense disambiguation both in terms of the problems caused for acquisition (Ribas, 1995a; Resnik, 1993a; Li & Abe, 1998) and for application to the task (Federici et al., 1999; Resnik, 1997; Ribas, 1995a). In the next chapter, we characterise the field and explore some methods that might be helpful in tagging the data input to the preference acquisition.

2.4.4 Handling of Proper Nouns

As well as being hampered by the lack of sense disambiguation on the input data, previous research has pretty much ignored the issue of proper nouns. In early work, Resnik (1992) divided proper nouns equally between the WordNet classes **location** and **someone**. These two classes, along with **organization** cover the majority of proper nouns. They feature as roots in the hyponym hierarchy and have clear semantic differences. In later work, Resnik (1993b, 1993a) mapped all proper nouns to **someone**. Other researchers have avoided proper nouns altogether. This is perhaps a reasonable stance in face of the considerable ambiguity they pose. However, proper nouns make up a sizeable proportion of the argument heads of noun phrases. By employing some software for identifying proper nouns, more data should be covered and the additional data will be less ambiguous. The extent to which the new data is disambiguated, and correctly disambiguated, will depend on the coverage and accuracy of the proper noun identification process.

Identifying numerical quantities such as dates, time periods and money is relatively straightforward and can be done with pattern matching techniques. Proper noun recognition, on the other hand, requires a lot more consideration. General Architecture for Text Engineering (GATE) (Cunningham et al., 1995) is a software environment with information extraction components freely available to those doing research. The named entity recogniser and classifier makes use of large lists for identifying well known organizations (mainly companies), locations (chiefly cities and countries) and people (common names). In addition, trigger words such as *ministry* or *airlines* are used. Finally, a proper name grammar is used alongside a bottom-up chart parser to detect and classify multi-word proper names.

A portion of the BNC, providing 1.8 million words of parsed text, was run through the GATE named entity recognition component. [12] On account of the time taken for processing, three days on a SUN sparc Ultra, we did not process any further sections of text. [13]

The SCF lexicon produced using the 1.8 million words (lexicon B) contained 17049 proper names which could not be classified. Additionally 11073 were classified as **person**, 3986 as **organization** and 2790 as **location**. We experimented with the data for the object slot and found this increased the number of argument heads by 18%. The large quantity of unclassified proper nouns do not leave things any worse off than before since they are simply ignored. There are however some obvious sources of error. The proper noun *Wexford* occurred frequently in the data and 161 occurrences were classified as **location**. From manual inspection, it seemed that quite a number of these should have been classified as **person**. One such example was:

(8) I am going to take a shower, Wexford said coldly.

[12] This was the VIE NE recognition system in version 1.1.

[13] There were additional problems encountered when running GATE, related to the size of the files and long sentences.

Earlier in the text, there is a section on *Wexford House* followed by a lengthy section relating to a *Chief Inspector Wexford*. There is an additional module of a larger information extraction system from the same team that might help with this. The module is called the discourse module and is described by Wakao et al. (1996). The co-reference component of this module should help with some of these errors. Semantic inferences such as selectional preferences, might also improve the accuracy. For example, in 8 above, it is **people** or **organization**s that tend to act as the subject of *speak*. In the GATE system, these semantic inferences are only applied to unclassified proper names and are not used to change classifications already given.

We conducted an experiment using lexicon B. A sample of 28 verbs were selected at random, subject to the constraint that they exhibited multiple complementation patterns and that they occurred with more than 20 argument heads at the direct object slot in lexicon B. The ATCMs at the direct object slot were compared with and without the proper noun resolution. The results indicated that proper noun resolution does substantially improve coverage. Out of 28 verbs only 7 were cut at the root as opposed to 12 without proper noun resolution.

We leave formal evaluation for discussion in chapter 4. From informal evaluation, there are many verbs for which the ATCMs with and without proper noun recognition are similar. However, for some verbs there are striking differences. For example, without proper noun recognition, the ATCM for the direct object slot of *move* is cut at the root. Whereas with proper noun resolution, the ATCM includes a preference at **location** (association score 1.02). The benefit of proper noun recognition will vary depending on the verb and slot combination. The effect of this will also depend on the sublanguage of the corpus. Intuitively, one would expect Wall Street Journal processing to benefit substantially from correct identification of organisations. The benefits to be had have to be weighed against the additional cost. For the bulk of our experiments, and for diathesis alternation prediction, we abandoned proper noun recognition because of the heavy processing load it placed on our system.

2.4.5 The DAG Issue

One further problem with the structure of WordNet arises because WordNet is actually a DAG rather than a tree. This was acknowledged by Li & Abe. They accounted for this by splitting the hierarchy into subtrees at each case of multiple-parents in the search downwards for the best cut model. They therefore duplicated the nodes beneath the multiple-parents, effectively creating new senses. They were not explicit about how the frequency counts were shared between the duplicated nodes. In our implementation, the frequency counts are incremented to ancestor classes only once for each descendant. In areas where there are multiple parents, the frequency contribution from a child to multiple parents is its full frequency count, and this is propagated to all parents. At these layers in WordNet, this means that the sum of the probabilities of classes on the cut will exceed one. We think this is intuitive in that if the parent classes jointly cover offspring then the sum of the probabilities of the classes at the level of the multiple parents reflects this overlap.

Thus, in the example shown in figure 2.10, a frequency count of 1 detected at the **person** node will only be incremented once at the superordinate **entity** class. A cut at this class will therefore meet the probability axioms. A cut at the layer of **causal agent** and **life form** would however give rise to a sum of the probabilities on the cut greater than 1. This is understandable when we take

Table 2.6: Frequency by depth of classes with multiple-inheritance

Depth	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Freq	0	0	1	3	22	111	161	147	68	21	12	11	0	1

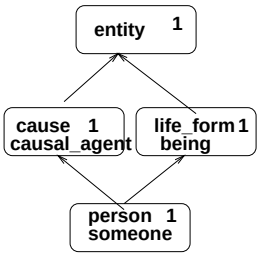

Figure 2.10: An example of multiple-inheritance

into account that multiple-inheritance gives rise to overlap between the two parent classes both in terms of the subsumed nouns senses and their probability.

Aside from the probabilities not summing to one at levels involving multiple-parents, there is another unresolved problem. This arises because the method used to obtain the most appropriate cut across WordNet relies on a tree structure and not a DAG. In the search for the best cut we, like Li & Abe, follow all paths from a shared parent. But, in our scheme, shared descendants are not given new senses. This may mean that there are classes in the resultant cut which are related by hyponymy. We remove duplicates so that if the two paths suggest the same class then this is added to the cut only once. It is, however, possible to append two cuts which overlap because one path suggests a higher level than the other. Removing nodes on the cut which overlap in this way can result in gaps in the tree cut and is a costly enterprise, we therefore rejected such a practise. Like Li & Abe we do not circumvent the problem. We have not attempted to compensate for this fact because of the small extent of the problem. On examination of the cases of multiple inheritance, we find that less than 1% of the classes in the noun hyponym network have more than one parent class. Moreover, the majority of such classes are deep down in the network, as can be seen from table 2.6, and many concern compound nouns, for example *school boy* and *head nurse*, which are not handled by our system. For ATCMs the vast majority of these cases are classes which are pruned by virtue of having a probability beneath the specified threshold.

Cuts at the level of overlapping classes (the parents of shared offspring) will be penalised because of the additional classes, and penalised or promoted because of the extra probability in this area. For the ATCMs and LLRTCMs, if high preference scores are involved at this layer, then the layer may be given preferential treatment because of this additional probability. For the LLRTCMs, the converse will occur in areas of dispreference. The PTCMs will be penalised at layers of overlapping classes.

2.4.6 Acquiring Preferences Specific to SCF

Since we use a SCF lexicon as a starting point, we are able to extract data specific to the SCF, as well as to the verb and slot. This is important for collecting the data required for diathesis detection. Additionally it is interesting to see the extent to which collecting data specific to SCF affects selectional preference acquisition. Collecting data specific to a SCF may help by alleviating some of the noise brought about because the arguments in the different SCFs are carrying out different underlying semantic roles. For example, collecting direct objects [14] from any frame will allow *money* (the 'theme') from example 9(a) to be amalgamated with *woman* (the 'recipient') from 9b).

(9) a. He gave money.

b. He gave the woman money.

Taking data specific to the SCF will not of course remove all of the noise. This is illustrated in example 10, where both (a) and (b) would be assigned the same SCF, although the argument heads are carrying out different roles.

(10) a. He pays the man.

b. He pays money.

The drawback to using SCF specific data, is the reduction in available data. For example, the direct object slot for the `[np vp np]` frame, contains 37% less argument heads compared to the number of argument heads from all the frames which have a NP directly after the verb. The reduction in data will be considerably large for rarer frames. The reduction in noise, which compensates for the reduction in data, will vary considerably depending on the actual verb. Looking at the direct object slot of *give* when the data is obtained from all frames, the PTCM indicates a conditional probability at the **person_1** node of 0.29, **possession** on the other hand is 0.02 and **object** 0.07. When the transitive frame is considered in isolation, **person_1** is given only 0.07 and classes which more typically express the 'theme' of *give*, **possession** and **object**, are slightly higher than before (0.03 and 0.09 respectively).

2.5 Summary

The focus of this chapter is a survey of the techniques and choices for selectional preference acquisition. Following a review of the associated merits and weaknesses of the different approaches, we adopt one originally proposed by Li & Abe (1995). In this class-based approach, selectional preferences are generalised by WordNet classes. The correct set of classes are found using the MDL principle which finds the set of classes that makes the best compromise between being a good fit for the data and providing a succinct model. This method is applied to argument head data enumerated within an automatically acquired SCF lexicon, so that data can be obtained specific to a slot, and if required to a SCF. Modifications are suggested to the original approach. These modification include adding leaves for all internal nodes. This allows the approach to be applied to the full structure of WordNet whilst coping with a larger input data set. An option is provided for

[14] We use the terms direct object and indirect object as 'surface' syntactic labels. In this thesis they bear no relation to any notion of underlying semantic role.

using a preference score better equipped for low frequency items. Ambiguity in the input data is an important problem. In this chapter we looked into preprocessing the proper nouns with an existing system. In the following chapter we move on to ways of handling ambiguity of the common nouns which form the bulk of our input data.

Chapter 3

Word Sense Disambiguation for Selectional Preference Acquisition

---

WSD has long been associated with selectional preferences. The relationship is somewhat circular. Preferences are semantic in nature and should ideally be acquired from sense tagged data. The chief obstacle is that there are no sense tagged resources large enough for full scale acquisition. Corpus data contains word forms and not senses, and it is the latter that are required for inferring selectional preferences. The other side to this circular relationship is that selectional preferences can be used for WSD (Federici et al., 1997, 1999; Resnik, 1997; Ribas, 1995a). The semantic preference of a predicate for its arguments can help to determine the correct semantic class for a particular argument instance.

Preference acquisition has been performed, for the large part, on untagged data. One exception to this was Ribas's experimentation (1995a) using the 200,000 word portion of the Brown corpus sense tagged under SemCor (Miller et al., 1993a). This portion is also available parsed in the Penn Treebank II corpus. Ribas pointed out that acquisition from ambiguous data results in a considerable degree of noise in the resulting preferences. His experiment was performed as a way of indicating the benefits of using sense tagged data: the benefits were demonstrated compared to acquisition from the same quantity of untagged data. However, the limit on the quantity of data available has an adverse affect on the preferences obtained, with or without disambiguation. The costs of producing sufficient hand tagged data for full scale acquisition are prohibitive. Even if the manpower was available for large scale semantic tagging, this would defeat the major advantage of automatic acquisition, that it can be applied to a new corpus without significant overheads.

A second exception was the system of Pozanski & Sanfilippo (1996). They acquired semantically annotated SCFs using the information within the Longman Lexicon of Contemporary English (LLOCE) (McCarthur, 1981) to achieve semantic disambiguation and labelling. The grammar codes provided in LLOCE were used to help identify SCF tokens extracted from the corpus. LLOCE semantic codes which did not conflict with the syntactic evidence were then assigned to the SCF tokens. This research demonstrated how the knowledge within MRDs, and corpus evidence can be usefully combined. The approach depended on the availability of an MRD with the required information. No formal evaluation was undertaken of the acquired information.

Basili et al. (1993) also acquired selectional preferences from semantically annotated data, however, they manually annotated the argument head data. This is a costly enterprise which becomes increasingly expensive as the corpus size increases. Manual tagging has generally been avoided for automatic lexical acquisition because of the high level of human effort required. This chapter explores ways in which sense tagged argument head data could be produced automatically, on the assumption that selectional preference acquisition is better performed on disambiguated data, and taking into account the performance and requirements of current WSD techniques.

In the following section we look at the specific requirements of disambiguating word senses for preference acquisition. In section 3.2, we briefly summarise current techniques in WSD. In section 3.3, we identify three techniques that are taken forward for experimentation. In the subsequent section, 3.4, the three options are explored to investigate their performance on nouns in the SemCor data and other, randomly selected, nominal data. Two options are chosen in section 3.5 for disambiguating the data used for preference acquisition. Section 3.6 demonstrates the differences observed in the preference models when these options are applied.

3.1 Requirements

3.1.1 The Targets

We require disambiguation of the corpus data input to the selectional preference acquisition system described in chapter 2. This data consists of verb tokens each with the nominal argument head occurring in a specified syntactic relationship with the verb token. This data can be viewed as tuples of the form <predicate, slot, argument head>. Before extraction of the tuples from the corpus, there is an opportunity to disambiguate the data with recourse to the full context. Otherwise, any disambiguation will need to be made with only the information contained in the tuple. Either option is possible.

Although the verb and argument head could both be targets for disambiguation, in this thesis we restrict ourselves to disambiguating the argument heads only. This decision was made, in part, because the SCF which we use lists entries by verb form and not sense. The contextual information that might separate verb senses is no longer present when we look at the lexicon. Furthermore, if we generate sense specific preferences, they cannot readily be placed back in the lexicon which lists entries by verb form. We could, however, have generated a lexicon specific to verb sense by tagging verb forms before building the lexicon. There was a reason for not doing so.

The main reason for not pursuing disambiguation of the verbal predicates is the prerequisite for a sense inventory. We would require a distinction between related and unrelated senses to enable us to amalgamate the data from related senses to alleviate sparse data problems. Furthermore, fine grained distinctions would complicate matters for diathesis alternation identification. Diathesis alternations are often accompanied by a slight change in meaning. They are relevant to verb senses rather than verb forms, since it is the meaning of the predicate that gives rise to the syntactic behaviour. The meaning components of a verb which give rise to the syntactic behaviour are hard to pin down. Specification of the senses of a verb which should be collapsed for diathesis detection would presuppose some of the lexical knowledge which we are endeavouring to acquire automatically.

A wise move would be to distinguish only broad senses of a verb. The traditional distinction

between 'homonymy' and 'polysemy' might prove useful in this case. Traditionally, the term 'homonyms' is used where different words have the same form. The shared word form is coincidental. One example is *ear*, where the **organ of hearing** sense and the **part of a cereal plant** sense have different etymologies, that is they come from distinct roots historically (Lyons, 1977). Polysemy is reserved for cases where there are related senses of the same underlying word. A significant hurdle to an approach which relied on this distinction is that manually devised sense inventories do not always adhere to it.

There is considerable scope for the automatic production of sense inventories for a particular task. Brown, Pietra, Pietra, & Mercer (1991) devised such a technique for machine translation. Sense distinctions were only considered where they gave rise to different translations. There has also been research producing sense inventories from a training corpus (Schütze, 1992, 1998), using automatic clustering of contexts. Whilst these results are encouraging, automatic clustering can require manual editing before application.

In our approach, the selectional preferences of each verb and slot combination are represented in the TCMs as a set of classes across WordNet with associated preference scores. Different senses of any particular verb typically give rise to differences in argument structure and preferences, unless the senses are closely related. Some slots will, unfortunately, be common to a number of senses. For example, the object slot of *serve* is common to the **set at table** sense and to the **be useful to** sense. However, the polysemy of a verb is brought out by the profile of preference scores along the cut model. The argument heads occurring with both senses will contribute to the TCM. Thus, for example in the ATCM on the cut for the direct object slot of *serve*, illustrated in figure 3.1, the preference at the **substance** class reflects the sense of serving food or drink to a person, meanwhile that at **group** reflects work done for an organisation.

The argument head data used for diathesis alternation acquisition is specific to SCF. This reduces the number of verb senses involved for a given verb, slot and SCF combination. For example, both the **sack** and **shoot** senses of *fire* have a subject slot, but only the **shoot** sense has the intransitive frame. However, there is a residual problem that diathesis alternations involve two (and sometimes more) frames. Because of this, it is quite possible that several different senses of a verb will contribute data used to obtain selectional preference models for a combination of two SCFs. Moreover, it is quite possible that the two target frames under scrutiny will not both involve the same set of senses. For example, the transitive frame of *fire* arises from both the **sack** and **shoot** senses. This is illustrated in the examples 11(a) and (b) below. Meanwhile the intransitive frame is only permitted with the **shoot** sense, as shown by 11(c). 11(d) is semantically anomalous with *fire* in the **sack** sense. Verb polysemy will undoubtedly make it harder to detect alternations in some cases.

(11) a. The chief fired the gun.

b. The chief fired the boy.

c. the gun fired.

d. *the boy fired.

Although we avoid semantic disambiguation of verbs, we experiment with disambiguation

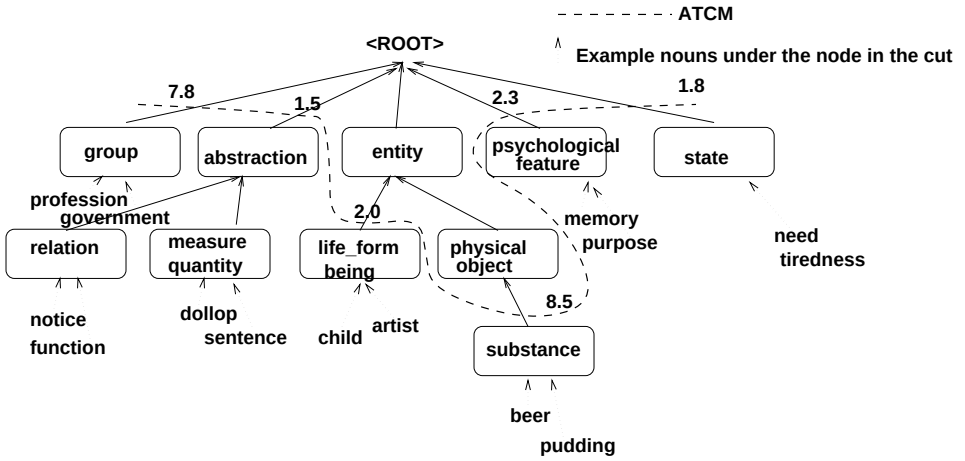

Figure 3.1: Serve direct object slot

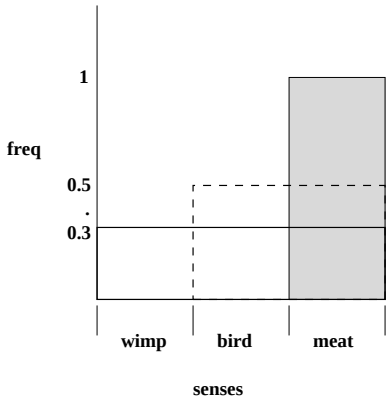

Figure 3.2: Assigning frequency credit to alternate senses

of the nominal argument heads. For this experimentation, we require a WSD method capable of tagging each argument head with the relevant WordNet class (or classes).

The original preference acquisition system copes with ambiguous data by dividing the frequency credit from each lemma by the number of senses of that lemma (see page 24 in chapter 2). For example, *chicken* has three senses in WordNet: the **meat** sense, the **bird** sense and the **wimp** sense. If *chicken* is observed, then $p(class)$ for each of these senses would be $\frac{1}{3}$. Thus, the frequency credit is spread uniformly between all 3 senses. This situation is shown in figure 3.2 by the white rectangle with a solid outline. Disambiguation aims to associate the total frequency credit with only one of these classes. This is illustrated by the shaded rectangle. It is not essential that all argument heads be disambiguated. Any that cannot be handled can be left ambiguous and treated as before using the uniform distribution over the senses of the lemma. It is also possible, given our scheme, to allow partial disambiguation, removing some senses but not necessarily leaving one. The remaining senses can share the frequency credit between them. For example, a system might identify that a token was either the **meat** sense or the **bird** sense, but definitely not the **wimp** sense. This is shown in figure 3.2 by the white rectangle with a dashed outline. Whilst we do not require full disambiguation, we do aim to tackle a large proportion of the ambiguity.

3.1.2 Evaluation and Accuracy

Human sense tagging is an extremely costly enterprise (Resnik, 1997). It is not an option if we want an automatic method of preference acquisition that can readily be applied to new material without additional overheads. We require an automatic WSD system that can automatically tag the majority of nouns that appear in argument head position.

One important fact to bear in mind when comparing WSD systems is that many results reported in the literature are evaluated in terms of a small sample of words (Wilks & Stevenson, 1998b). Performance is hard to compare across systems since the results of any system will vary considerably depending on the sense distinctions of the target words. Some systems may fare better on different types of sense distinction. Results from a small sample of words, even if the same set is used by different systems, may not give a clear indication of how well things would work on a large sample.

Evaluation is a notoriously difficult area. As well as differences in the target sample, the actual test and training data used will affect the results. Some results reported in the literature have included monosemous words (Ribas, 1995a; Wilks & Stevenson, 1998b), for which one can only expect 100% accuracy. It is important to have some sort of baseline provided alongside the test data which indicates the degree of difficulty of the task. Usually the random baseline is used. This is the result one would get if a random selection was made in every case. This is calculated by taking the reciprocal of the number of senses for each test item, and averaging over all these items. This is shown below in equation 3.1 where the summation is over k test words (w). In cases where systems use supervised training data, it is common to cite the baseline obtained by choosing the most frequent sense in every case. This is usually referred to as the 'first' or 'predominant' sense. The first sense baseline is usually higher than the random baseline.

$$\sum_{i=1}^{k} \frac{\frac{1}{|senses_{w_i}|}}{k} \times 100 \tag{3.1}$$

Other aspects of the training and test data are also important. Some systems are able to exploit syntactic or semantic preprocessing to good advantage whilst others are able to handle raw text. The requirements of different systems make the results harder to interpret. The results of different systems should be compared with the coverage and requirements of the systems in mind.

In the light of these difficulties in evaluation, the SENSEVAL (Kilgarriff et al., 1998; Kilgarriff & Palmer, 2000) competition and workshop was devised in an attempt to establish a level playing field for WSD systems to make comparison easier. Unfortunately the organisers had to drop the task of tagging all words in a sample (the 'all words task') because of lack of resources. It is precisely algorithms that can handle 'all nouns', or a significant proportion of them that we require. Systems that work well with a handful of words may not perform so well on a different test set, or the all words task (Wilks & Stevenson, 1998b). Nevertheless the SENSEVAL competition provided a useful means to comparing the merits of systems on a target set of 35 words.

WSD tasks were set up for each of the test words. For some tasks, the senses involved all belonged to the same POS of the target word. The POS was not supplied for other tasks. Participants could attempt any subset of tasks, or the full set. Systems that required supervised training data were given the same training set. No restrictions were made on the quantity or nature of data for

unsupervised training. A crude three way classification of systems was given so that systems could be compared alongside others in the same group with similar training requirements. The three way classification was supervised (S), unsupervised (A) and requiring other training (O).

The sense distinctions from HECTOR (Atkins, 1993), a corpus lexicography project, were used with tasks at three levels of granularity. The three levels were:

1. coarse grained - the main HECTOR sense divisions
2. fine grained - respecting all subdivisions of the main HECTOR senses
3. mixed grained - system responses were given full credit if they were more specific than the manually tagged item, and partial credit if they were more general.

HECTOR was used because the inventory has not been cited in the WSD literature and so none of the participant systems would be at an unfair advantage. All systems that relied on another inventory were faced with the task of mapping to HECTOR senses for the competition. In addition to the English competition, a parallel competition (ROMANSEVAL) was run for researchers using Italian.

The level of accuracy of the WSD is obviously important, but is not critical for selectional preference acquisition. This is because we are not considering lemmas in isolation, but collectively in groups bearing a specified relationship with a predicate. A system requiring results for each target word individually, such as a machine translation system, would not tolerate too many mistakes. For preference acquisition we combine the disambiguated data and feed this into a statistical process. Mistakes, though undesirable, simply add to the noise. The level of accuracy should of course be as high as possible. Minimally, preference acquisition should fare better with than without the WSD. We can compromise a little on accuracy, but there are other factors that we need to consider for tagging the majority of nominal argument heads. What we require is a system that can produce a reasonable level of accuracy whilst coping with large volumes of data.

3.1.3 Machine Processing Time and Human Effort

It is important to establish the overheads involved when using WSD systems on a large data set. Some algorithms require a substantial quantity of tagged data for supervised training. Manually producing sense tagged data is a costly enterprise. There is a small amount of sense tagged data available in existing resources, for example SemCor and the DSO corpus (Ng & Lee, 1996). The quantity of data in SemCor (200,000 words) is too small for many supervised WSD systems. The DSO corpus, meanwhile, has only a specific set of target words which are disambiguated. For some applications there are alternative sources of training material. For example, Brown et al. (1991) make use of parallel text for WSD for machine translation. The utility of supervised systems is limited to situations where resources, such as aligned corpora or sense tagged data, are available.

As well as the cost of producing any data required, there are machine processing costs to consider. Large training overheads may not be a problem for a small handful of test words, or a small quantity of test data. However, the training overheads will quickly mount up for the all words task as the quantity of material for disambiguation increases. Performance at run time is also important.

3.1.4 A Digression on Precision and Recall

Evaluation of WSD systems is frequently performed in terms of *precision* and *recall*. These measures are also used for reporting results in lexical acquisition, but there are subtle differences between the way that the measures are calculated for lexical acquisition and the calculations for WSD. The differences are naturally related to differences in the task, but they affect the way in which the two measures relate to one another. In this chapter, we report results of WSD experiments, however, much of this thesis concerns lexical acquisition. It is therefore important to point out the differences in the calculations, and we do so here, before we discuss any results reported in the literature, or report our own results in this thesis.

For lexical acquisition, the task involves identification of tokens as positive or negative occurrences of a particular phenomenon. The phenomenon under observation might be a particular SCF or a selectional preference for a particular semantic class. The counts that are usually collected for evaluation are:-

- true positives (TPs) - types (or tokens) recorded in the gold standard which are correctly identified by the system
- false positives (FPs) - types (or tokens) which are incorrectly identified by the system. They do not actually occur in the gold standard.
- false negatives (FNs) - types (or tokens) in the gold standard which are not identified by the system

Precision is the proportion of guesses that the system makes which are correct.

$$Precision = \frac{number\ of\ \text{TPs}}{number\ of\ \text{TPs} + number\ of\ \text{FPs}} \tag{3.2}$$

Recall is the proportion of items in the gold standard which the system guesses correctly.

$$Recall = \frac{number\ of\ \text{TPs}}{number\ of\ \text{TPs} + number\ of\ \text{FNs}} \tag{3.3}$$

There is a often a compromise between obtaining a high precision and a high recall. One can make lots of poor guesses and achieve a high recall by covering many of the items in the test set but, at the same time, one will obtain a low precision since many guesses will be wrong. A high precision can be achieved by taking a more conservative stance, and only guessing when one is really sure.

True negatives (TNs) are not usually taken into account since there will typically be a large number of phenomena which are correctly not identified in a given context. One could calculate the percentage of things that were correct (or wrong):-

$$Accuracy = \frac{number\ of\ \text{TPs} + number\ of\ \text{TNs}}{number\ of\ \text{FPs} + number\ of\ \text{FNs}} \tag{3.4}$$

However, the number of true negatives is typically large, and, as Manning & Schütze (1999) point out:

> One can get extremely high accuracy results by simply selecting nothing (Manning & Schütze, 1999, pg.269)

Thus recall and precision measures are often used for evaluation.

Recall, precision and accuracy are also measures used for evaluation of WSD systems. However, the task is somewhat different to lexical acquisition and the differences are reflected in the calculations. A verbal description of the measures does not bring out these differences, and so the differences are not obvious. Accuracy is again the number of items labelled correctly, over the total number of items. Recall is the proportion of items in the test set which are correctly labelled, and precision is the proportion of items which the system labelled correctly. The difference in the use of these measures for WSD, compared to their use in lexical acquisition, is that the set of items which the system labels, the denominator in precision, is a subset of the items in the test set, the denominator in recall. Precision can only be greater, or the same level as recall. The system is not asked to make guesses on items for which a label has not been specified. The positive-negative distinction is not really relevant and the tradeoff between precision and recall is not quite the same. Moreover, for WSD, accuracy is equivalent to recall. The term accuracy is usually used when the system makes a decision on all test items and so there is no distinction between precision and recall.

$$Precision = \frac{number\ of\ correct\ assignments}{total\ number\ of\ assignments} \tag{3.5}$$

$$Recall(Accuracy) = \frac{number\ of\ correct\ assignments}{total\ number\ of\ test\ items} \tag{3.6}$$

Recall and precision values reported in this chapter are in respect of WSD experiments, and the values are calculated as in equations 3.5 and 3.6. In subsequent chapters, precision, recall and accuracy figures are calculated as described in equations 3.2, 3.3 and 3.4 unless otherwise stated.

3.1.5 Summary of Requirements

To summarise, disambiguation of the argument head data requires a system which will:-

1. tag the majority of nouns with one or more senses from WordNet
2. be reasonably accurate
3. make little or no use of manually produced data, and
4. have acceptable training time and run time requirements

The next section provides a brief overview of the WSD literature. We do not attempt to describe all WSD systems, since there are so many. We will instead provide a broad categorization of systems and discuss some of the noteworthy systems. We will also single out systems which might be suitable for our purposes.

3.2 Background

The field of WSD is vast. A full history and survey of the field is beyond the scope of this thesis, but we can refer the interested reader to the literature review provided by Ide & Véronis (1998). The aim of this section is to give a characterisation of the classes of WSD systems and the requirements and merits of these. Our classification is a modified version of the one provided by Charniak (1993). Systems are classified according to their use of a priori knowledge (manmade MRDs and machine readable theasauri (MRTs) and corpora (with and without sense tagging). To illustrate our classification, we describe some of the systems that are widely cited in the field. There are four main groups in our classification:

1. Knowledge-based approaches
2. Statistical approaches using external knowledge
3. Supervised statistical approaches
4. Unsupervised statistical approaches

In the first approach, a priori knowledge is used to select the appropriate sense given the target context. The a priori knowledge might come from MRDs or MRTs or manmade rules. In the second approach, corpus statistics are collected for the entities contained in a manmade resource, such as a MRD or MRT. The manmade resource is used to structure the collection of statistics, and in some cases to produce an automatically tagged corpus for training. The training data helps tailor the manmade resource to the target data. In the third approach, statistics are collected from a corpus of sense tagged data. Finally, in the fourth approach, the statistics are collected from training data without recourse to prior knowledge or sense tagged data.

3.2.1 Knowledge-Based Approaches

These approaches make use of prior knowledge and the context of the target word. Manmade heuristic rules with domain specific contextual cues can be used for disambiguation. Many researchers have used the knowledge which exists in dictionary definitions to circumvent the overhead of knowledge acquisition from an expert (Lesk, 1986; Cowie, Guthrie, & Guthrie, 1992; Veronis & Ide, 1990; Agirre & Rigau, 1996; Wilks & Stevenson, 1998b).

Cowie et al. (1992) used definitions from LDOCE (Procter, 1978). They adapted Lesk's (1986) original idea of using the overlap between the dictionary definitions of the target word and those of the other words in context. Cowie et al. took all the words to be disambiguated in a sentence together, instead of concentrating on one target word at a time. The first sense in LDOCE was selected for each of the words in the sentence as a first approximation to the solution. For each word sense in this initial combination, the words from the dictionary definition were located and stemmed. This is illustrated in figure 3.3 with the target sentence *The butcher usually skins the chicken*. For the diagram we have simplified and removed some of the sense entries from LDOCE. The lemmas from the definitions of the first sense of each content word are shown next to that word, with a dashed underline. The overlap of these lemmas from the definitions was used in a 'redundancy measure'. This redundancy measure provided an indication of how cohesive the combination of senses was. The system was then faced with the task of searching for the optimum

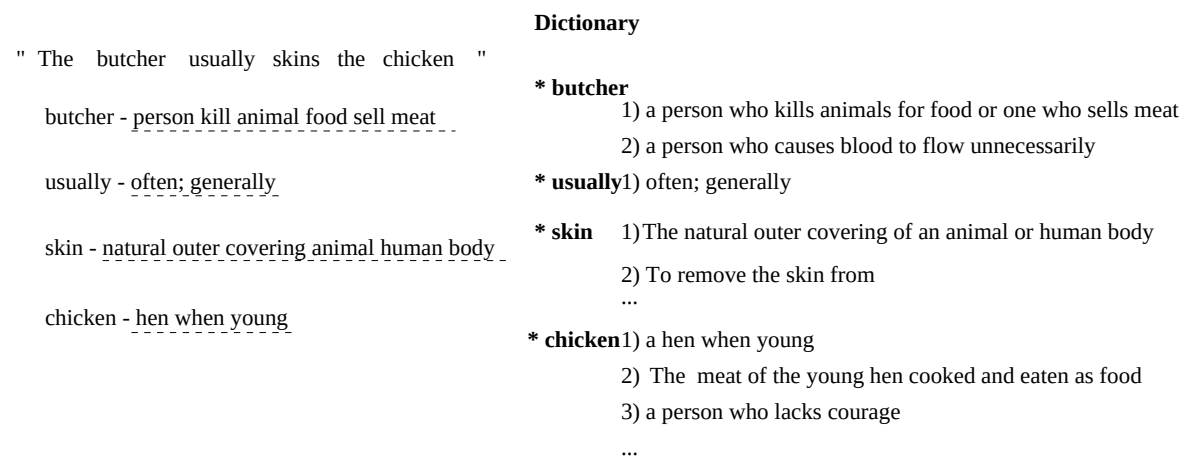

Figure 3.3: Using dictionary definitions for the content words in a sentence

combination of word senses over the entire set of possibilities for the sentence. The search was conducted using 'simulated annealing', a technique for solving combinatorial optimisation problems. The term 'simulated annealing' is used in an analogous manner to the way in which metals cool and anneal. In the WSD system of (Cowie et al., 1992), the redundancy measure was used within an 'energy' score which at each iteration guided the replacement of one sense from the current configuration. Results were reported for a sample of 50 sentences at 47% accuracy to the LDOCE sense level, and 72% to the LDOCE homograph level. There were 5.5 ambiguous words per sentence, on average. The authors were not explicit about the time taken for the disambiguation process but state that this was reasonable. This approach has strong appeal because it covers all words in the sentence concurrently. By doing so, it allows the disambiguation process to draw evidence from the disambiguation of all the words collectively.

Veronis & Ide (1990) also utilised the definitions from a MRD. They used definitions in the Collin's English Dictionary (Hanks, 1979) to make connections between word forms and word senses which were represented as nodes in a neural network. The word nodes were positively linked to the nodes which represented all the possible senses of that word. The sense nodes had positive connections to word nodes where the words existed in the definitions of that sense. Negative connections were made from each sense to the other senses of that word, so that activation at any particular sense for a word diminished the activation at competing senses. Disambiguation was performed by activating all the nodes representing words in the target sentence. As in the work of Cowie et al. (1992), all words in a sentence were disambiguated together. Activation then spread through the network iteratively. Finally, each word from the sentence was assigned the sense with the highest activation value. Véronis & Ide did not provide a quantitative evaluation.

Wilks & Stevenson (1998b) reported results from a WSD system which combined a variety of knowledge sources to perform sense tagging of all content words with LDOCE sense tags. POS tagging was used before other knowledge sources were applied, however the POS tagging was overturned if it did not accord with the senses given in the dictionary. The knowledge sources comprised the pragmatic codes and selectional restrictions provided in LDOCE and the dictionary definitions of LDOCE which were used in a modified version of the Cowie et al. algorithm. De-

cision lists were used to combine these knowledge sources. The decision lists were inferred from supervised training data by a machine learning algorithm. Results were reported on a 2021 word subset of the SemCor corpus. This subset was automatically tagged with LDOCE tags by using a mapping between WordNet and LDOCE senses. 1821 words were used as training data and results were reported on the remaining set of 200 words. Accuracy was reported at 92% by projecting the results to those expected for the overall corpus, assuming 100% accuracy for monosemous words. From these results, Wilks & Stevenson concluded that a high level of accuracy could be attained when working on the all words task, as opposed to a small sample of words. This was certainly a useful finding, albeit on rather a small test set.

Agirre & Rigau (1996) presented a system for tagging all nouns in text using the information held within WordNet. One noun was taken at a time. The candidate senses of this noun were found in WordNet, along with the senses of the other nouns occurring within a context window. The window was specified as a fixed number of nouns, with the target noun in the centre of the window. Non-noun words were not used for disambiguation. A subhierarchy was identified for each sense which included that sense, but did not overlap with the subhierarchy of another sense. 'Conceptual density' scores were calculated for the subhierarchies that the candidate senses belonged to. The conceptual density measure contrasted, for each subhierarchy, the number of senses from the window context belonging to this subhierarchy, with the number of senses within the subhierarchy in the first place. The candidate sense belonging to the subhierarchy with the highest conceptual density was selected. Results were reported from disambiguation of all polysemous nouns in 9,000 words of the SemCor text. Precision was 43% and recall was 34% to the WordNet sense level, given a window size of 30 nouns. Agirre & Rigau advocated the use of conceptual density alongside other knowledge sources, rather than in isolation.

Results reported for knowledge-based methods are promising, particularly in systems which combine multiple knowledge sources (Wilks & Stevenson, 1998b). Also, much of the research using these methods has been evaluated on the all words task. This has been possible because the reliance on manmade knowledge avoids the requirement for sense tagged data or heavy training overheads. A significant drawback to these approaches is the heavy reliance on handcrafted information. This may not be a problem if the manmade resource matches well with the target text. However, performance may be affected when the technique is transferred to a different corpus. In our preference acquisition we did make use of the manmade resource WordNet. However, we did so with recourse to corpus statistics. WordNet presented us with a way of structuring our statistical models. Since we wished our preference models to take corpus data into account we preferred WSD approaches that did likewise. There are approaches to WSD that make use of the combination of a priori knowledge and corpus data and it is to these we now turn.

3.2.2 Statistical Approaches with External Knowledge

The WSD approaches in this category use corpus statistics collected with regard to the entities contained in manmade resources. These approaches still rely on manmade knowledge, but this reliance is reduced. The impact on WSD of any inadequacies of the manmade resource are reduced by the statistics. Frequency counts from the corpus support the a priori knowledge only where it is relevant to the data. There is, unfortunately, no compensation when portions of the data are not

reflected in the resource. This can happen where there are sense omissions. Using large resources, such as WordNet, keeps this problem to a minimum.

One such approach is the application of automatically acquired preferences to WSD (Resnik, 1997; Ribas, 1995a; Carroll & McCarthy, 2000). Preferences are acquired from untagged data and then applied to the disambiguation task. This is done by selecting the sense of the target noun with the highest preference score given the verbal predicate and slot.

Resnik (1997) tested the preferences for 100 strongly selecting verbs on the data in SemCor. He used his selectional association score to predict the sense of objects, subjects, nouns in PPs and nouns in head–modifier relationships. He achieved overall accuracy of 44%, averaged over these relationships, with a random baseline of 29%.

Ribas (1995a) performed two experiments. In the first, he used only 4 mid-frequency verbs *rise, report, seek* and *present*. For these he obtained a precision of 80% compared to a test set of hand tagged examples, with a random baseline of 63%. Additionally he tried his preferences on the full set of argument heads within SemCor. He obtained a disappointing 53% for these, which was not significantly different from the random baseline of 52%. Ribas explained the difference between his two experiments with regard to differences in the quantity of data for training, corpus differences (the WSJ having less sense distinctions then the BROWN), and the selectional properties of the verbs involved.

The substantial differences in Ribas's and Resnik's results can be attributed, at least in part, to differences in the test samples. The random baselines in Ribas's experiments were high when compared with the random baselines in Resnik's. This was because Ribas included monosemous nouns in the sample. The variation in baseline brings home the importance of evaluating on the same test data.

The SENSEVAL competition included two participant systems using selectional preferences alone. The results are publically available (Rosenzweig, 1998). One of these was the SUSSEX system (Carroll & McCarthy, 2000), which used selectional preference models (ATCMs) produced using the system described in chapter 2. For the all nouns task, fine grained precision to HECTOR word senses was 40.8%. The other system was the OTTOWA system (Kilgarriff & Rosenzweig, 2000). This system obtained a precision of 33% for the same all nouns. The random baseline on this task was 30% with a phrase filter to handle the easy multi-word cases (14.6% without).

Looking at these results, it is evident that selectional preferences are not a panacea for WSD. However, they do provide a useful source of information for systems which combine evidence, such as that proposed by Wilks & Stevenson (1998b). They used selectional preferences supplied in the LDOCE dictionary.

There are other ways of mixing prior knowledge with statistics. Yarowsky (1992, 1995) developed two WSD approaches which used external knowledge to automatically sense tag data. The sense tagged data was then used for supervised training.

In (1992), Yarowsky disambiguated words to one of the 1024 categories in Roget's thesaurus (Chapman, 1977). To do this, he collected a sample of representative contexts for each of the categories. He used concordances from a corpus (Grolier's Encyclopedia) around words belonging to the category. For example, *shovel* belongs to the category **tool**. Occurrences of *shovel* in the corpus contributed to the representative contexts of **tool**, along with occurrences of other

members of this category. Yarowsky then identified the salient words for each of these categories by collecting statistics over the words in the representative contexts. These statistics were then used to identify the correct category for a target word. For example, *crane* might belong to **tool** or **animal**. He achieved 92% accuracy to the level of the Roget category on the 12 words tested.

The second approach (Yarowsky, 1995) was described as 'unsupervised' WSD. This approach only required external knowledge in the initial stages. It relied on initial seed collocations to discriminate senses that could be observed in a portion of the training data. This portion was labelled accordingly. New collocations were then extracted from the labelled sample and ordered by a *log-likelihood ratio*. This measure is given in equation 3.7, where *sense A* of a word is distinguished from the *other senses* of the same word. This measure should not be confused with the *log-likelihood ratio* tests discussed in the previous chapter, although the same terminology is used.

$$\text{log-likelihood ratio} = \log \frac{p(sense\ A|collocation_i)}{p(other\ senses|collocation_i)} \tag{3.7}$$

In Yarowsky's (1995) system, the new ordered list of collocations was used to relabel the data. The system then iterated between observing and ordering new collocations, and relabelling the data, until the algorithm converged and the residual untagged data was stable between iterations. The final decision list of collocations was then applied at run-time.

Although the approach hinged on the initial seed collocates, these only needed to provide initial tagging for a small proportion of the training data, typically between 85 and 98%. Furthermore, the seed collocations did not have to be very accurate since poor collocations were weeded out as training proceeded. Yarowsky produced the seeds manually but suggested that they might be found from MRDs or on-line resources such as WordNet. Yarowsky reported results on 12 words each with two possible senses trained on a 460 word corpus. He reported 96% accuracy for this sample of test words.

3.2.3 Supervised Statistical Methods

In a supervised training approach, a substantial portion of sense tagged training material is required for estimation of the parameters. These parameters are the contextual clues that co-occur with individual senses. Statistics are collected for these contextual clues in the text surrounding each particular sense.

The sense tagged data can sometimes be produced without manual sense tagging. Brown et al. (1991) used bilingually aligned corpora, within the context of a machine translation application. The meaning of a word was determined by its translation. For example, in French to English translation, *prendre* is translated as both *make* and *take*. This was taken to indicate two different senses of *prendre*. Mutual information scores were collected for words co-occurring with the different translations. These context words were then used as disambiguators on new candidates for translation. Evaluation was performed in the context of machine translation. For a small set of 100 sentences, translation was improved from 37% to 45% by the WSD component. This evaluation does not make for easy comparison with other work in the WSD literature. However, it did demonstrate WSD improving performance on a real task; something that is often assumed but not proved.

In situations where bilingual data is not available, or not relevant, one is faced with the prospect of manually tagging data. Yarowsky (1994, 1993) applied supervised methods to tasks where the tagged training material was already available. In (Yarowsky, 1994), supervised WSD was applied to accent restoration. In this experiment, French accents were placed into text where they were previously omitted. This was done by collecting statistics from other texts containing the accents. The contexts which distinguish accents were identified and then used for restoring accents in the target data.

In (1993), Yarowsky investigated the hypothesis that collocations are reliable indicators of sense. Again, he used supervised methods; collecting statistics for words in the context of specific senses. A variety of sense distinctions were taken to make use of already existing sense tagged material. Yarowsky used distinctions from translations (e.g. *prendre - make/take*), homophones (e.g. *cellar/seller*), words that might be confused in optical character recognition (e.g. *terse/tense*), traditional hand tagged material as well as "pseudo-words" where two words (e.g. *covered/waved*) were artificially combined for experimentation purposes. Pseudo-word experimentation was performed by replacing each occurrence of either word with a combined form e.g. *covered-waved* in the data for testing. Corpus data with the words in their original form was then available for training to determine the relevant collocations which distinguished the individual words (*covered* and *waved*). For some of these sense distinction options there were legitimate applications, with sense distinctions relevant to the task. For other types of sense distinction, notably pseudo-words, the technique simply enabled experimentation with a disambiguation procedure in the absence of suitably hand tagged material.

The analogy-based approach of Federici et al. (1999, 2000) was a supervised approach to WSD, implemented for Italian. We have previously described this approach on page 18 in chapter 2 since it relates to the automatic acquisition of selectional preferences. Here we see that the acquired selectional preferences can be applied to WSD, just as other WSD approaches make use of selectional preferences. As we saw in section 2.2.2 on page 18, this approach exploited an example-base which stored instances of predicate-argument relationships. One such example was *fumare-sigarette/Object*, which gave the English equivalent *smoke-cigarette/Object*). These stored instances were organised in analogical families. When words, for example *fumare* and *accendere* (to light), shared a particular relationship (predicate-object) with another word (*sigarette*), then a link was made between them, and they were placed in the same analogical family. Inferences were then made for a novel predicate-argument instance, such as *accendere-pipa/Object* (light-pipe), if an instance from the same analogical family was already stored in the example-base with the new argument, such as *fumare-pipa/Object*. WSD was possible because a portion of the example-base was sense tagged. Examples were linked not only because they shared common contexts, but also because the senses of predicates and arguments were preserved. Candidate senses for a target instance were compared with regard to tagged examples of the target predicate stored in the example-base. The rival candidate senses were ranked according to a weighted measure of the number of contexts supporting this sense, and the semantic entropy of the 'pivot' terms (*fumare-sigarette/Object* in the above example). The semantic entropy took into account the number of collocate types. This allowed evidence from words with more specific sets of collocates to be given more weight. On the Italian equivalent of SENSEVAL, ROMANSEVAL, the system performed

at an impressive 81% precision.

The results in this section are promising. However, in the absence of sufficient training material, supervised training is likely to be uneconomic for most NLP applications.

3.2.4 Unsupervised Statistical Methods

Unsupervised approaches do not rely on human input. The unsupervised WSD systems in the literature use distributional similarity as a means of classifying, and disambiguating data.

Schütze (1992) proposed such an approach using the representation of semantic space that we described on page 17 above. In this scheme, co-occurrence data from raw text produced a vector space in N dimensions, where N was the number of context words used for mapping. Words which appeared within a fixed window around a word were used to represent the position of that word in semantic space. The distance was measured in characters (1000 or 1200, this was varied for experimentation) to allow for the fact that more informative words tend to be longer. [1] Word types were represented by combining the contextual evidence from all occurrences of the word. Word tokens were represented using the normalised average (centroid) of the vectors for the words in the context of the occurrence. To identify senses for disambiguation, a training set of untagged occurrences of the target word was used. Vectors representing the contexts of these occurrences were clustered. The clusters were then manually assigned the relevant senses. WSD involved identifying which relevant cluster (sense) was appropriate for a novel target context. Rather than using the frequency distribution of the words within the window of the target word directly, Schütze took the centroid of the vectors already plotted for the words within the context window. This was then compared to the cluster centroids representing the distinct senses using the cosine between two vectors as a measure of similarity.

An example of this approach, adapted from one in (Schütze, 1998), is given in figure 3.4. For clarity, this example only considers two dimensions, for the context words *clothes* and *court*. In this example, the word *suit* is taken to have only two meanings, a **legal** sense and a **garment** sense. In this example, the training set of occurrences of *suit* results in two clusters, C1 and C2. C1 is identified as the **legal** sense, and is closer to a vector for the word *witness* and the *court* dimension. C2 represents the **garment** sense and is closer to the *clothes* axis and a vector representing *laundry*. A new instance of *suit*, requiring disambiguation, is plotted according to the centroid vector of its context words. The centroid vector represents the target context. The closest sense (C1 or C2) to the newly plotted target is selected. Again, similarity is measured using the cosine between two vectors. Note that with this method there may be more clusters than senses resulting from the automatic clustering process and that the labelling of these clusters is performed manually. Assigning a label to the classes is only required for WSD where sense tagging for a pre-existing inventory is required, for example, for finding the correct translation in machine translation, or the correct pronunciation in a text-to-speech system. Sense labelling is also required for evaluation against a gold standard data set. The clusters representing senses could be used without labels in a system where recourse to a pre-existing sense inventory is not required, for example in information retrieval (Schütze & Pederson, 1995).

To obtain a realistic semantic space Schütze (1992) used thousands of context words from a

[1] Later, in (1998), Schütze used a window of 25 words either side of the target word.

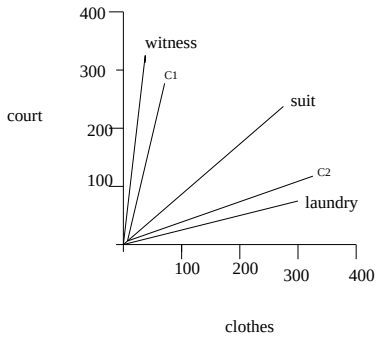

Figure 3.4: Schütze's disambiguation without outside knowledge

corpus of 50 million. To overcome the performance overhead of storing and processing all this data Schütze relied on singular value decomposition, which is a form of dimensionality reduction that finds the principal axes in vector space. Schütze (1992) reported results for ten test words each with a binary sense distinction, except one which had a three way distinction. There were between 100 and 500 test instances for each test word, and on average the system obtained an accuracy of 92%.

3.3 Selecting Candidate WSD Approaches

We wished to explore a few WSD techniques from the multitude available for disambiguating nominal argument head data. The high scores of some of the systems are alluring but comparisons between systems are hard to make because of differences in the evaluation task. This is especially the case where only a small number of words have been tested. The actual sense distinctions taken make a large difference to performance (Leacock, Towell, & Voorhees, 1993). For selectional preference acquisition, there were other considerations to be made, aside from accuracy. We needed to examine the requirements of the system in terms of the resources required. These resources included the quantity and nature of any training data, training time, and the amount of human-effort involved. As suggested in section 3.1, these were particularly important for the all nouns task required here.

Since our goal was to automatically acquire preferences directly from corpus data we sought approaches to WSD which did likewise. Approaches which directly apply external sources of knowledge show encouraging results (Cowie et al., 1992; Wilks & Stevenson, 1998b). However, they rely on the human endeavours that have gone into building these resources. This ties the systems to the resources, and the disambiguation process cannot readily be tailored to reflect the peculiarities of the corpus at hand. Furthermore, many such resources are not freely available.

The absence of human input makes unsupervised approaches rather attractive. They would naturally go hand in hand with an approach to acquire preferences without reference to a manmade resource. However, in chapter 2 we outlined our reasons for choosing WordNet instead. The unsupervised approaches characterise similarity solely on the basis of distributional evidence. This can give rise to incongruous classes. Using external knowledge to help constrain the collection of statistics is shown to be a promising approach. When Yarowsky (1995) compared his results with Schütze (1992) on the same four test words, Yarowsky's unsupervised algorithm performed

at 96.7% against Schütze's 92.2%, even in the face of a lower baseline (55% as opposed to 65% in Schütze's experiments), and therefore a harder task.

Supervised approaches are problematic for our task. There is no corpus of sense tagged data of sufficient size to estimate the parameters for all argument nouns, or even a substantial portion of them. Since we wished to avoid both (i) the quirks of distributional data and (ii) total reliance on manmade resources, using statistical methods with external knowledge was the obvious choice. The preference acquisition process outlined in the previous chapter made use of a training corpus (the written portion of the BNC) and WordNet. These resources were also used in our WSD experiments.

The results reported in section 3.2.2 above indicate that selectional preferences are not a complete solution to lexical ambiguity. Selectional preferences do, however, improve performance over the random baseline. The acquisition of selectional preferences is a subgoal of this thesis. As a consequence, exploiting these selectional preferences provided an obvious choice for reducing the ambiguity of the input data, in an iterative approach. In this approach, selectional preferences were first obtained on fully ambiguous data. These preferences were then used to filter erroneous senses from the input data. The refined input data was then used for a new cycle of selectional preference acquisition. Experiments using automatically acquired preferences for WSD experimentation are reported in section 3.4.1.

Yarowsky's 'unsupervised' WSD method (1995) showed particularly good results for the handful of words it was applied to. It also had the advantage that it only relied on a small amount of handcrafted knowledge for the initial seed collocates. We explored this method for the all nouns task, since this method avoided the need for supervised data, and only a small amount of prior knowledge was needed.

Whilst avoiding knowledge-based and supervised approaches, we did experiment with a third option which was related to these. This option was that of using the first sense of a word, regardless of context. This heuristic has been used as a lower bound baseline by many researchers (Yarowsky, 1995; Schütze, 1992; Gale, Church, & Yarowsky, 1992). However, this is not appropriate for systems which do not require tagged data, since supervised data or prior knowledge is required to obtain a ranking of senses in the first place (Resnik, 1997).

Wilks & Stevenson (1998a) have indicated that this heuristic can produce good results to the homograph level when used in conjunction with POS tagging. They used the ranking of senses provided in LDOCE to define the first sense for each word form. Wilks & Stevenson used POS information together with this heuristic on all words in a 1700 word corpus from the WSJ. They reported results at 92% accuracy. The test data included monosemous words. Wilks & Stevenson (1998b) went on to use this heuristic in conjunction with other sources of information from LDOCE. This WSD system was described above in section 3.2.1 on page 55.

Gale et al. (1992) used the first sense heuristic to characterise a lower bound for evaluation. In their experiments, they used a random sample of 97 words, which included 30 polysemous words. When unambiguous words were included, they obtained a baseline of 93%, using the first sense heuristic for the polysemous words. When polysemous words were used exclusively, the first sense baseline was 81%. There was considerable variability when using this heuristic. For some words e.g. *virus* (2 senses) the distribution was skewed and the first sense baseline was high,

(98%). The predominant sense was less prominent for other words, for example, *output* (2 senses) had a low baseline (51%). WSD systems, and particularly when tested on words like *virus*, do not always outperform this simple heuristic (Gale et al., 1992; Rosenzweig, 1998).

This heuristic produces surprisingly good results if it is compared to a baseline selecting one of the senses at random. For all words, a first sense heuristic will produce better results than the random baseline if the sample size is large enough. The improvement will be larger for words with highly skewed distributions such as *virus*. The first sense heuristic is straightforward to apply, which makes it a straightforward WSD candidate for experimentation on the all nouns task. SemCor (Miller et al., 1993b) provides a portion of WordNet tagged data from which we determined the first sense of nouns occurring in both our data and the SemCor data. Indeed the senses in WordNet were ordered according to this data, where the senses have been found in SemCor. [2] We experimented with this WSD heuristic on the all nouns task and report our findings in section 3.4.2.

3.4 WSD Experiments

In this section we present results on lexical disambiguation for the three WSD options selected in the last section. The three options are:

1. using selectional preferences
2. the first sense heuristic
3. Yarowsky's 'unsupervised' technique

We evaluated these approaches both in terms of their performance on the all nouns task and in terms of the overheads that they would place on the selectional preference acquisition process. Evaluation of WSD systems is far from straightforward, as one can see by the wide variety of ways of comparing systems in SENSEVAL (Rosenzweig, 1998). For evaluation, we used the data in SemCor. All words are labelled in this corpus. The DSO corpus also has WordNet labels, but in this corpus only a selected sample of words are labelled. SemCor is freely available and for this reason it has been used for evaluation by many other researchers (Ribas, 1995a; Agirre & Rigau, 1996; Resnik, 1997). Also, Wilks & Stevenson (1998b) map LDOCE sense tags into WordNet senses in order to evaluate with SemCor.

For evaluation of the first sense heuristic, our training and test data were equivalent since we obtained the sense ranking to identify the first sense from SemCor. On account of this overlap, we also tested the heuristic on small manually tagged samples of randomly collected data from the Lancaster-Oslo/Bergen (LOB) (Johnansson, Leech, & Goodluck, 1978) and WSJ corpus.

3.4.1 WSD Using Preferences

The method of acquiring selectional preferences was described in the previous chapter. To recap, the acquired preferences, for a specific verb and slot, were acquired as a set of disjoint classes across WordNet covering all leaves. All word senses were attached at leaves in our modified version of WordNet. The preferences were found by populating the hierarchy with frequencies from the corpus data. MDL was used to obtain a set of classes at an appropriate level of generalisation.

[2] Otherwise the ordering is random. For this reason the SemCor data was required, and the ranking within WordNet could not be relied upon.

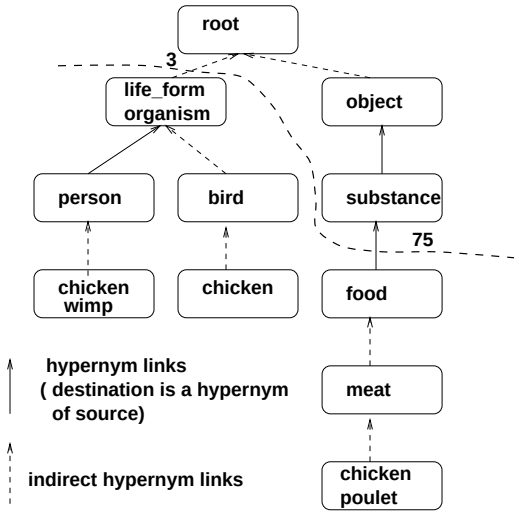

Figure 3.5: Direct object slot *eat*

Using preferences for disambiguation was straightforward. The classes on the cut with candidate senses underneath were compared. The sense, or senses, selected were those under the class with the highest preference score. For example, the sense of *chicken* under **food** would have been preferred over the senses under **life form**, when occurring as the direct object of *eat*, given the selectional preferences in figure 3.5. The disambiguation process acted as a filter, removing senses only where they fell under different classes on the cut. Disambiguation could be quite coarse grained depending on the specificity of the cut and the semantic proximity of the senses. For example both classes containing *curry*, (**curry powder** and **curry dish**) lie under **food**. Both would have been selected with this method given the tuple $< eat, direct\ object, curry >$.

ATCMs were used for WSD evaluation. A threshold was used with the association score. Disambiguation was only performed if at least one of the classes had an association score above the threshold. The threshold was used because it is hard to be sure of a negative correlation using scores based on mutual information (Church et al., 1991). Filtering senses, rather than returning one solution, did not pose a problem for selectional preference acquisition. As we saw in section 3.1.1, partial disambiguation was acceptable. If more than one sense was returned the frequency count was simply split between these senses.

Although our selectional preference acquisition system could readily handle cases where more than one sense was returned from the WSD component, we had to allow for this possibility in our evaluation. The gold standard (SemCor) which we used for evaluation has, for the large part, one sense label to each lemma. Where the WSD system returned multiple labels, we scored the target lemma correct if at least one of the assigned senses was correct. We permitted this because erroneous senses were at least close, in WordNet terms, to the correct sense. This was because the system only selected multiple senses which were hyponyms of the same superordinate class on the cut with the highest preference score, as in the $< eat, direct\ object, curry >$ example above.

It was easier to score highly under this scheme since we were effectively allowing the system more guesses. We were increasing the chance of selecting the correct sense for each item by the number of senses left after WSD. To reflect this we provided a new baseline since the random base-

line was too low in these cases. The random baseline (RBL) is given in equation 3.1 on page 50. An adapted baseline - the 'multiple choice random baseline' (MCBL) is given in equation 3.8. The summation was over the sample of (n) test items that we disambiguated. The random $\frac{1}{|senses_i|}$ for each candidate (i) was multiplied by the number of senses selected by our selectional preferences ($|choices_i|$).

$$\text{MCBL} = \sum_{i=1}^{n} \frac{\frac{|choices_i|}{|senses_i|}}{n} \times 100 \tag{3.8}$$

Since n was the number of items for which disambiguation was attempted, this was a precision baseline. We also calculated recall in our experiments. However, this can only be compared to the standard random recall. We could not calculate a multiple choice recall since for undisambiguated items we could not indicate additional weight for multiple sense assignments because the number of these (and therefore the granularity) were determined by the disambiguation process.

Resnik (1997) and Ribas (1995a) both made a random choice between the senses remaining after selectional preference disambiguation. We avoided this since humans often have difficulty choosing between senses (Kilgarriff, 1993). We prefer to think of the process as a filtering of unlikely senses. The SENSEVAL competition permitted systems to provide a probability distribution over the candidate senses of a test item. Systems that did not output a probability distribution used the uniform distribution over multiple assignments. Systems that identified only one sense ascribed all the probability to a single solution.

Our method of WSD permitted the selection of multiple senses. More than one choice was given only in cases where the senses fell under a common superordinate. Unfortunately, MCBL did not indicate the semantic proximity of the senses remaining after disambiguation. It was therefore rather high and punitive. A baseline between RBL and MCBL would have been more appropriate. Resnik & Yarowsky (1997) suggested incorporating a measure of semantic similarity in WSD evaluation to allow for multiple labels.

In addition to allowing multiple assignments, SENSEVAL compared systems on three levels of granularity (fine, coarse and mixed grain, as described above). This was an intuitive way of incorporating a notion of semantic similarity into the evaluation. Unfortunately, it only works with sense inventories for which a breakdown of senses into subsenses is provided. We did not do this for our SemCor evaluation since levels of granularity are less obvious in WordNet.

Evaluation was performed using the preferences collected from lexicon A for a sample of 30 verbs. These verbs were selected to exemplify a range of SCFs and were not chosen on the basis of their selectional properties. The verbs are listed below:

add, agree, allow, ask, begin, believe, bring, build, call, cause, change, charge, choose, consider, cut, decide, end, establish, expect, feel, find, fix, give, help, like, move, produce, provide, seem, swing

ATCMs were obtained for object, subject and PP slots. Disambiguation was performed on all nominal argument heads in these relationships. The results are provided in table 3.1. The results shown are with a threshold of 1 placed on the association scores, except in the case of 'obj 2',

Table 3.1: SemCor evaluation

slot	recall	recall RBL	precision	MCBL	RBL
obj	35	27	48	43	28
obj 2	24	27	46	37	25
subj	35	27	51	47	27
PP	9	25	27	40	26

where a threshold of 2 was taken. The 'recall RBL' column displayed the random baseline for recall.

The selectional preferences performed above both baselines in all slots except the PP slot. Performance at all slots was not significantly better than the rather stringent MCBL. [3] The results were highly significant when compared to the random baseline. Performance on the PP slot was affected by sparse data. The slot was less frequent than subject or object slots, even before we considered the specific preposition involved. The data available was substantially reduced when it was considered with respect to a particular preposition. Another possible reason for the poor performance at the PP slot might be that the head nouns in PPs are less constrained by the verb than the argument heads in other slots are.

Comparing the results with other researchers is not straightforward because of differences in the test data. These differences are indicated to some extent by differences in the random baselines. Ribas attained a score of 52% for verbs chosen randomly in the SemCor data (1995a). The random baseline was 51%. Monosemous words were included in the test data and so 52% is perhaps disappointing. The small size of the training data (the SemCor corpus) undoubtedly affected Ribas's results.

Resnik obtained 44% accuracy for the object slot, and 41% for the subject slot (1997). He did not perform the experiment with the PP slot. Where the preferences did not distinguish between senses, a random selection from the remaining senses was used. The random choice baseline was 29% for both slots. Resnik used data for 100 of the strongest selecting verbs, where selectional strength is defined as in equation 2.12 on page 27 in chapter 2.

In our SemCor experiment described above we used all verbs for which preferences were obtained. Using strongly selecting verbs would have produced better results since verbs with weak preferences are unlikely to be as good at distinguishing senses. In a preliminary experiment, we evaluated preference disambiguation on a manually tagged sample of the direct objects of *eat* from lexicon C (created from 1.8 million words of parsed text from the BNC). Only polysemous words were used. This provided a recall of 62% and precision of 93% (compared to MCBL of 55%). Performance was high because *eat* selects strongly for its direct objects.

Evaluation using the SENSEVAL test suite permitted comparison with other systems (Rosenzweig, 1998). The preferences we used were ATCMs, with the first sense heuristic for WSD of the input data. Our results varied substantially from 0% to 100% precision depending on the target word. Many errors were accounted for by factors other than the quality of the selectional prefer-

[3] The significance was tested using the χ^2 test (Siegel & Castellan, 1988)).

ences. These included multi-word identification errors, some of which were readily correctable, POS tagging errors, parser errors and Hector–WordNet mapping errors. Performance for nouns on the coarse grained task was 69.4% precision with 20% recall. This compared well with the OTTOWA system which also used only preferences. This system obtained 70.6% precision and 8% recall on this task.

We agree with Resnik (1997) that selectional preferences can only provide part of a solution to WSD. However, in our experiments they did outperform the random baseline and they provide a fully automatic approach without the need for sense tagged material. We would not recommend using them alone in applications where accuracy is critical. In section 3.6.1, we see whether they can improve selectional preference acquisition, where disambiguated data is used collectively.

3.4.2 Using the First Sense Heuristic

In this section we describe our evaluation of the first sense heuristic for diambiguating WordNet senses. The first sense of a given lemma was determined using the frequency information available from SemCor (Miller et al., 1993a). The success of this heuristic depends heavily on the frequency distribution over the senses (Gale et al., 1992). We therefore applied constraints on the application of this heuristic to improve precision. Where the constraints were not met the ambiguity was left unresolved and we reverted to the uniform distribution over senses for selectional preference acquisition.

Initially we experimented with two criteria for application of the heuristic.

1. FREQ - a threshold on the frequency of the first sense
2. RATIO - a threshold ratio between the frequency of the first sense and that of the next most frequent sense.

The first of these criteria ensured that there was a reasonable quantity of evidence collected for the first sense. The second ensured that the heuristic was applied only where the distribution over senses was sufficiently skewed.

Initially FREQ was set at 5 and RATIO at 2. The heuristic with these parameter settings was evaluated against SemCor. This is the corpus from which the frequency data was obtained so one would expect the results to be higher than where the training and test data are disjoint. The only other corpus tagged with WordNet senses is the DSO corpus (Ng & Lee, 1996). However, this is not free of charge, unlike SemCor which is freely available. To evaluate the heuristic on unseen data, we manually tagged two small samples from the LOB (179 nouns) and the WSJ (191 nouns). These samples were selected at random. The results are shown in table 3.2, along with the results from SemCor (Brown corpus). The SemCor results show performance on all nouns in the corpus.

Performance was understandably higher when scored against the same data from which the frequency ranking for the heuristic was taken. In all cases the heuristic outperformed the random recall (recall RBL). This was especially the case since we ignored cases that do not meet our constraints, as indicated by separate figures for precision. We included monosemous words for these experiments for ease of comparison with (Wilks & Stevenson, 1998a). They achieved an accuracy of 92% using the sense ranking provided in LDOCE. This figure was higher than our precision figures because Wilks & Stevenson disambiguated to the LDOCE homograph level, a considerably easier task.

Table 3.2: Threshold 5 ratio 2

DATA	(recall) RBL	recall	precision
Brown	45	61	86
LOB	35	44	69
WSJ	40	41	68

Table 3.3: Variation of thresholds on the LOB data

FREQ	RATIO	IDN	RECALL	PRECISION
5	2	N	44	69
3	2	N	47	69
1	2	N	49	67
3	1.5	N	50	67
3	3	N	39	76
3	2	Y	45	71

Further experimentation was performed using the LOB sample to find a good setting for the FREQ and RATIO thresholds. Additionally, a third boolean constraint was added: nouns identified on the SemCor project as being difficult for humans to tag were ignored. We use the term IDN (ignore difficult nouns) for this constraint. The frequency distribution estimated from the 'gold standard' corpus was likely to be flawed for these nouns because of the difficulties experienced by the human taggers. For this reason these nouns were not targeted for automatic tagging.

The results from varying the parameters are shown in table 3.3. The parameter settings of 3 for FREQ, 2 for RATIO and IDN were taken forward for selectional preference acquisition. These values were felt to make the best compromise between recall and precision from informal inspection. There is further scope for playing with these parameters but their effect on performance is not considerable. We were ultimately concerned with the effect of this WSD method on selectional preference acquisition. Further experimentation with these parameters was therefore only warranted if the method improved the selectional preferences acquired.

3.4.3 Yarowsky's Iterative Approach

We also investigated Yarowksy's (1995) unsupervised algorithm described in section 3.2. This has a distinct advantage over supervised algorithms in that it does not rely on manually tagged data. A little prior knowledge is required to provide seed collocations for initial labelling of a small portion of the training data. In our implementation, WordNet was used to generate the seed collocations to provide the initial tagged portion of training data. This was required for learning a first approximation to the decision list of ordered collocations.

Seed collocations were automatically obtained using the WordNet hyponym hierarchy. Seeds for a target noun were obtained in sets. One set was obtained for each class (sense) that the target noun belonged to. The seeds in a set for a particular class were taken from the set of synonyms at:

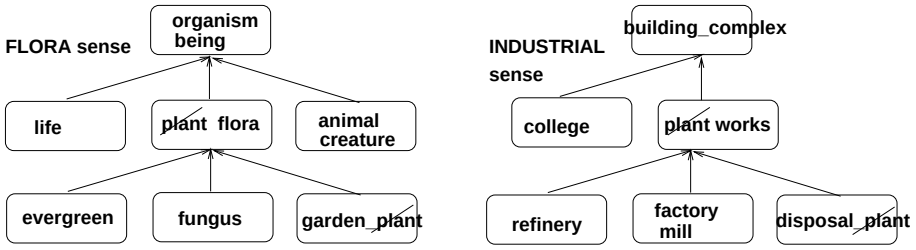

Figure 3.6: Seed collocates for *plant*

- the class itself
- direct parent classes
- sister classes
- child classes

Synonyms that featured as seeds in more than one set were removed from all sets for this target. As an example, some of the seeds for the two main senses of the target word *plant* are shown in figure 3.6. If the target noun occurred within multi-word expressions, then only words other than the target word were used as seeds. There are several examples of this in the diagram. For example, *garden plant* produces the seed *garden* for the **flora** sense, and *disposal plant* produces the seed *disposal* for the **industrial** sense.

Training data was collected from the written portion of the BNC, 90 million words approximately. This was unfortunately much smaller than Yarowsky's 460 million word corpus. In our experiments, the only type of collocation used is one within a fixed distance of words (10). This bag-of-words approach was simpler than using syntactic relationships. Performance was likely to be reduced, but the initial overhead of parsing the entire data set was avoided. we used a constant (0.1) in situations where a particular sense did not occur with a specified collocation, i.e. freq$(sense_A|collocation_X)$. Our algorithm stopped when more than 95% of the training data was tagged.

Our simplified implementation was initially evaluated on *plant*, which is the target given as an exemplar in Yarowsky's paper. We manually labelled 710 citations from the BNC with the appropriate WordNet senses. Only two WordNet senses were evident in the data (**flora** and **industrial**). On initial experimentation, it was evident that collocations involving senses with a higher frequency quickly overwhelmed the decision list, which was ordered by the log-likelihood ratio measure given in equation 3.7. This lead to a bias towards the predominant sense. In Yarowsky's example, sense A (**flora**) and B (**industrial**) of *plant* both had similar frequencies in the data labelled with the seeds. This would not be typical of corpus data and certainly was not the case for the data used here. Yarowsky did achieve satisfactory results with words having a clear predominant sense. Nevertheless, the log-likelihood ratio measure given in equation 3.7 favours the predominant sense since it is likely to have a higher conditional probability with respect to the collocation simply because it occurs more regardless of the context. To overcome this, the log of the

Table 3.4: Unsupervised WSD for *plant*

SCORE	RECALL	PRECISION
initial with seeds	16%	93%
log-likelihood	71%	72%
Association ratio	76%	78%

ratio of the association scores (see equation 3.9 and 3.10), rather than the conditional probabilities was investigated as a measure for ordering the decision list.

$$\text{log of the ratio of association scores} = \log \frac{A(targetsense, collocation_i)}{A(othersenses, collocation_i)} \tag{3.9}$$

$$\text{Where } A(sense, collocation_i) = \frac{prob(sense|collocation_i)}{prob(sense)} \tag{3.10}$$

3.9 can be simply rewritten for estimation as:-

$$\log \frac{\text{freq}(targetsense, collocation_i)}{\text{freq}(othersenses, collocation_i)} \times \frac{\text{freq}(othersenses)}{\text{freq}(targetsense)} \tag{3.11}$$

Where freq indicates a frequency count.

The results are displayed in table 3.4. Using Yarowsky's log-likelihood ratio as the score for ordering the decision list, the algorithm is biased towards the more frequent sense. Recall is 71% and precision 72% when the stopping condition is met. The ratio of association scores compensates for the relative frequencies of the senses. When the stopping condition is met in this case, the recall is 76% and precision is 78%.

In table 3.4 we also show the recall and precision obtained using only the seed collocations. This indicates how accurate our automatically generated seeds were. The seeds were only expected to cover a small portion of the training data, this explains their low recall value. However, the seeds for *plant* gave a precision of 93%. The automatically generated seeds were, on this occasion, very informative.

The association score ratio was better at ordering decision lists. The results reported here are far from the 95% accuracy that Yarowsky reported. However, in these experiments many simplifications were made which degraded performance. Better results can be expected with more training data, a more sophisticated method of smoothing, and parsing of the data to allow a wider variety of collocations. Despite our simplifications the results for *plant* were encouraging when compared to the random baseline of 50% (precision and recall).

Unfortunately, WSD on randomly selected targets involving finer word sense distinctions was not as successful. In a second experiment, training was performed for 391 polysemous mid-frequency nouns. The nouns were selected at random, except that the nouns identified on the SemCor project as being hard to tag by humans were excluded. Training was again performed

using the BNC data. We used the SemCor files of the Brown corpus for evaluation. The results were disappointing. We obtained 29% for both recall and precision, which was only just better than the random baseline (25%). An important factor was the poor quality of the initial seeds. When the initial seeds were used to label the test corpus, recall was 0.1% and precision was 18%. It is likely that performance could be improved for the all nouns task given more training data, more sophisticated smoothing, preprocessing to identify syntactically determined collocations and some refinement of the process to obtain seed collocations. Performance is unlikely to be at the 95% level with nouns involving less clear cut sense distinctions than the binary split for *plant*.

3.5 Choosing WSD Options

3.5.1 Preferences

Preferences are not a panacea for WSD. They may be useful, but complementary sources of knowledge are needed for accurate WSD (Wilks & Stevenson, 1998b). They did, however, present us with a method for disambiguation which outperformed random when applied to argument head data. We hoped that this would give improved results for selectional preference acquisition, compared to using the uniform distribution across all senses for each argument head. Since we had a method for producing the preferences this was an obvious choice for us. It did not require additional training, other than that which was necessary for selectional preference acquisition in the first place. The preferences were first approximated from ambiguous data and then acquired again from the partially disambiguated training data. We used a 'second pass' approach, with two preference acquisition cycles. The resultant preferences were compared to those produced using the uniform distribution over senses of each target before attempting further iterations. Disambiguation with selectional preferences required training for the verbs, rather than for the argument heads themselves.

3.5.2 The First Sense Heuristic

The first sense heuristic was also taken forward for preference acquisition. This was not because it is a good method for WSD; it clearly is not since it disregards context. Its chief advantage is in its ease of application. The training requirement was already met since we used the sense ranking provided by SemCor. This small corpus (200,000 words) was felt to be sufficient since we only had to estimate a ranking of the senses for each lemma. Additionally, we used our constraints (FREQ, RATIO and IDN) to apply the heuristic only in appropriate situations. Accuracy was not at the level that one would expect from state of the art WSD systems. However, accuracy was at a reasonable level, given that this heuristic had been evaluated on the all nouns task. Many of the state of the art systems have only been evaluated on a small sample of words. The accuracy level was not critical to us since the disambiguated argument head data was used collectively for selectional preference acquisition. The heuristic was intended to concentrate the data in the correct areas of preference, compared with use of the uniform distribution. It was hoped that the preference 'signal' would be stronger provided that the bulk of the data was assigned the correct sense. The erroneous senses would be scattered across WordNet, except in cases of highly frequent argument heads (Ribas, 1995a).

Table 3.5: Estimating training time for the all nouns task

freq threshold	noun types	polysemous noun types	training time (days)
0	104,912	22,978	239
10	18,424	7,071	74
100	5,593	3,684	38

3.5.3 Yarowsky's Algorithm

Yarowsky's unsupervised method undoubtedly worked well for some target words. It had a major advantage in that it did not rely on supervised material or extensive external knowledge. The algorithm, whilst initially taking knowledge derived from humans, settled on a decision list of collocates found statistically in the training data.

The poor performance of the algorithm for the 391 randomly selected nouns demonstrated that the nature of the sense distinctions, and the quality of the initial seed collocations, radically affected performance. Even if accuracy could be improved with refinements to the algorithm, syntactic processing of the training data and a larger training corpus, the time taken for training remained an issue. This was an important consideration for the all nouns task, depending on the size of the target corpus. We estimated training time figures for the nouns in Lexicon A from the training time required for *plant*. [4] Our algorithm took 15 minutes on a SUN Ultra before 95% of the data was tagged.

Table 3.5 displays the training times we estimated for the nouns types in the corpus data used to build lexicon A. There are approximately 22,978 polysemous (according to WordNet) noun types. We estimated that this would take nearly eight months of training. To reduce this, one option was to concentrate WSD on the most frequent items, since these cover a larger quantity of the data according to Zipf's law. The fourth column of table 3.5 shows the training time estimate for the nouns whose frequency exceeds the threshold specified in the first column. The nouns with a frequency more than 100 would require more than five weeks of training. 94% of the argument data could potentially be covered if we performed disambiguation with only these nouns. A significant problem with concentrating WSD on the highly frequent nouns is that they tend to be more polysemous and therefore harder targets for disambiguation. A more sophisticated implementation of the algorithm and a substantial increase in the quantity of training data would be required to handle these nouns. A larger training corpus should improve accuracy but would bring additional costs in terms of training time.

Yarowsky's approach was not taken forward for tagging the nominal argument heads. This was because of the poor performance on the 391 randomly selected nouns and the substantial training requirements. This approach worked well with coarser sense inventories and may indeed be useful for other WSD applications, provided that good sources of seed collocations can be found.

In the next section we look at the effects on the TCMs of WSD of the argument head data using the selectional preferences and the first sense heuristic.

[4] We acknowledge that the training time will vary depending on the actual word.

3.6 Preference Acquisition From Partially Disambiguated Data

In this section, we contrast the TCMs with disambiguation of the argument head data using (i) selectional preferences and (ii) the first sense heuristic. Formal evaluation is deferred until chapter 4 but we make some informal comparisons of the cuts produced with and without these WSD strategies.

Lexical disambiguation should have the effect of reducing noise from erroneous senses. If no disambiguation is performed, the frequency credits are divided evenly between all senses of each lemma, correct and incorrect. Because many lemmas are considered together, the credit will tend to concentrate in areas of preference. However, there will be some noise in other areas from parser errors and semantically 'odd' tuples as well as from erroneous senses. If disambiguation is performed, and if this is reasonably accurate, then the level of noise should be reduced. The frequency credit will increase in the areas of preference and decrease in other areas. The argument head data translated to WordNet senses will be more homogeneous, that is it will be more concentrated in areas of preference. The TCMs will consequently display more marked preferences than TCMs acquired without WSD. On the other hand, if the disambiguation is inaccurate, the frequency counts will be more widely dispersed than without disambiguation. This is on the assumption that the erroneous senses (those not genuinely associated with the verb) from different lemmas are independent.

For example, in an imaginary corpus, suppose that the lemmas *room, house, wall* are observed at the direct object slot of *decorate*. The different senses of the respective lemmas (indicated by the lemma and a suffix for the sense number) fall under the WordNet roots shown in figure 3.7. If no disambiguation occurs, the frequency count at the respective roots is as shown in the diagram in the row labelled NO WSD. The senses selected by a poor WSD algorithm are shown within dashed lines. This algorithm incorrectly labels the majority of argument heads and spreads the frequency counts from all lemmas throughout the hierarchy. The frequency count at the roots is shown in the row labelled Bad WSD. The use of the first sense heuristic selects the senses encircled with the solid line. The frequency distribution at the roots is labelled FirstS in the diagram. The first senses of these words all congregate under the same **entity** root in this case. This gives us a strong indication of the preference for *decorate*. In a real corpus we would have far more data. Some of the data would be monosemous. Some of the polysemous words would be like *wall*, where all senses are in the same vicinity of our semantic classification. Other polysemous words will have senses in different areas. Disambiguation will have most effect for these words with senses in different areas.

Verbs differ in their selectional properties. These differences can be seen in the profile across the TCMs in terms of:

1. the magnitude of the preference score
2. the specificity of the preferences
3. the dispersal of the preferences throughout the hierarchy

On the whole, WSD should increase preference strength in the appropriate regions. For some verbs there will also be an increase in the specificity of the TCMs. More specific cuts are typical in

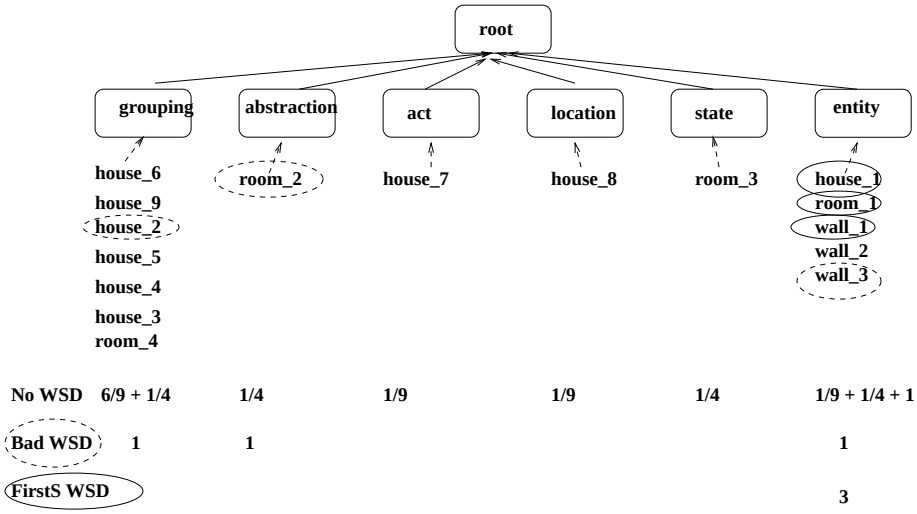

Figure 3.7: WSD and estimation of frequency distributions

regions with a large differential between the frequencies of subclasses. The frequency credit needs to be sufficiently high for this to happen. Additional frequency credit makes it more cost effective for MDL to take a deeper cut. The additional cost of a more detailed model is compensated for by the reduced data description length.

In the following sections, the effect of WSD on the TCMs is demonstrated in a number of ways. In section 3.6.1 we illustrate the differences by displaying portions of some TCMs with and without disambiguation. These illustrations demonstrate the effects of WSD qualitatively. Formal evaluation is deferred until the next chapter, but we also make some quantitative comparisons in this chapter. In section 3.6.2 we show the difference WSD made to the number of cuts at the root, showing the effect WSD had on the number of verbs for which TCMs were obtained below the dummy root. This provides us with a quantitative comparison in terms of the specificity of the TCMs.

The degree to which the argument head data was more homogeneous (more concentrated in areas of preference) with WSD is demonstrated in section 3.6.3. This is done by displaying some conditional probability ($p(c|v)$) distributions across a cut at the WordNet roots [5] for a couple of verbs with and without WSD.

There is a further way of quantifying the degree to which the disambiguation reduces noise and increases the homogeneity of the argument head data. This is inherent in our use of MDL to obtain the TCMs. The cost of the model, given the data, is used to guide the selection of the best model. An increase in homogeneity typically decreases the cost of a model. When using the same data, and the same method of calculating the relative cost (i.e. the method for ATCM, PTCM or LLRTCM), we saw if the disambiguation decreased the cost associated with the best model found. We make this comparison in section 3.6.4.

The disambiguation options in all sections below are indicated by NOWSD (no disambiguation), SPass (using the selectional preferences), FirstS (using the first sense heuristic), and COMB (a combination of the first sense heuristic and the preferences). For the combined option the first

[5] These are the eleven top level classes in the WordNet noun hyponym hierarchy

sense heuristic was used and where this could not be used the selectional preferences were applied.

The results below are reported on the acquisition of preferences for verbs found within the random sample of 500 sentences mentioned first on page 36. The data used was obtained from lexicon A.

3.6.1 TCMs

This section provides a qualitative description of the effect of WSD on the TCMs. We illustrate the differences by showing portions of the cut models for a few verbs. This demonstrates how well the preferences accord with intuition. We illustrate the differences using:

1. *produce* - a mid-frequency verb with preferences in a number of areas.
2. *melt* - a low frequency verb with strong preferences in one area.
3. *slice* - a low frequency verb where, without disambiguation, MDL selected a TCM at the root.

Figure 3.8 shows the ATCMs for the direct object slot of *produce*. SPass and NOWSD both cut the hierarchy at the **object** class. In this case, SPass weakened the preference score because stronger preferences elsewhere on the cut attracted the frequencies of ambiguous nouns. For example, *target* had five senses at hyponyms of the classes **location**, **lifeform**, **object**, **psychological feature** and **relation**. There was a strong preference (4.2) for *produce* at **relation**. This preference disambiguated the noun *target* occurring as direct object to *produce*, and removed the frequency credit under **object**.

FirstS tended to give more detailed cuts compared with SPass. This was because FirstS disambiguation resulted in one specific sense label for a given argument head. In the case of *produce*, FirstS provided a model which distinguished a stronger preference for **product** (*book, software* etc...) than for **construction** (*house, office* etc...). However, when we applied a combination of FirstS and SPass the resulting cut fell beneath the **product** class, we would perhaps wish for a higher level of generalisation than this.

In contrast *melt* had a lower frequency (36 compared to 2223 for *produce*). It had a prominent preference in the vicinity of the **substance** class. The ATCMs for the verb *melt* involved the same classes, by and large, regardless of the different WSD options. There were, however, differences in the association scores. Figure 3.9 displays part of the ATCM for *melt*. This portion was at the same classes for all WSD options. The association scores were different and are listed separately in table 3.6.

From this table we can see that all options provided a strong preference for the class **substance**. SPass further increased the homogeneity of the data when used on a verb with strong selectional properties. It typically provided the highest preference scores with strongly selecting verbs. FirstS helped where preferences were less strong, or where preferences were spread in different parts of the hierarchy, as with the *produce* example given above. Formal evaluation of these WSD methods are required over a large set of verbs before we make any judgement over the relative merits of these WSD options.

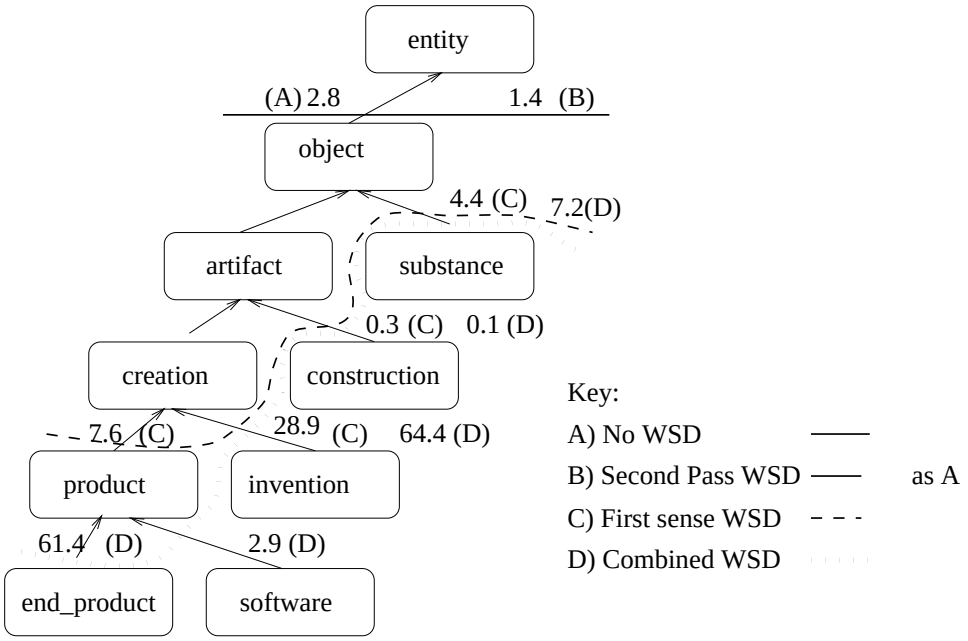

Figure 3.8: Models for *produce* object slot using different WSD strategies

FirstS had advantages over SPass in cases where a cut at the root is obtained for the NOWSD option. In these cases, the SPass technique could not be applied to the verb specific data. [6] *Slice* was one case where FirstS enabled reasonable preferences to be obtained. Figure 3.10 illustrates the cut model obtained for both FirstS and COMB. The model accords with intuition. The preference for **location** is less intuitive, but is readily explained by the arguments *top*, *middle* and *place*.

Although erroneous senses from different lemmas should be spread in different areas, erroneous senses from multiple occurrences of the same lemma will accumulate. Ribas (1995a) noted the accumulation of erroneous senses from lemmas which occur frequently with a verb. If a lemma is frequent with respect to a verb and slot, then disambiguation will concentrate this frequency even more. This will be for the better or worse, depending on the accuracy of the WSD. Frequent collocates were observed in our data e.g. *open door*, *slam door*. Where strong collocations occurred, cuts at the leaf classes were not unusual. This happened to a greater extent with FirstS than for SPass, since SPass at least took the verbal context into account. On the whole, FirstS correctly identified the sense of the lemma, as it did in *slam door*. Things did not always go well. For example, FirstS assigned *part*, in *play part*, to the **portion** class. Without FirstS, i.e. with either NOWSD or SPass, the frequency credit was split and so leaves were less common on the TCMs.

In this section, we used ATCMs at the direct object slot to demonstrate the effect of WSD. However, the effect also held for the other models (LLRTCMs and PTCMs) and slots (subject and PP). WSD tended to increase the specificity of TCMs, for all slots and model types. The specificity of different model types varied as we showed in the last chapter. Consequently, changes in specificity brought about by WSD for a particular verb did not always hold across different model

[6] For the LLRTCMs and ATCMs part of the prior data was disambiguated. These were the bits that co-occurred with verbs for which preferences were evident.

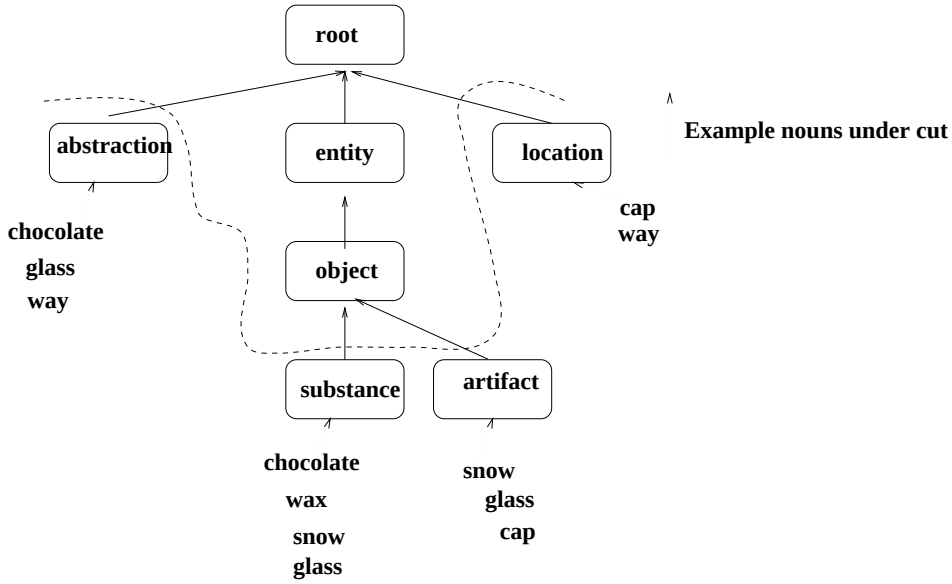

Figure 3.9: Classes on the ATCM for *melt* object slot using all WSD strategies

Table 3.6: Melt direct object preference scores for WSD options

	Abstraction	Substance	Artifact	Location
NOWSD	0.9	36.4	1.3	1.6
SPass	0.2	54.3	0	5.2
FirstS	1.0	38.6	1.3	0.3
COMB	0.4	54.1	0.5	0

types. PTCMs particularly tended to be less specific than ATCMs. Preference models at the PP slot suffered more from sparse data than subject and direct object slots. This was because the slot is less prevalent to start with, and also because the noun lemmas were considered with respect to the preposition as well as the verb. In the following section we use the percentage of root cuts across our sample of verbs to demonstrate the general tendency for increased specificity with WSD. We do this for the three different slots, model types and WSD options.

3.6.2 Percentage of Root Cuts

In this section, we list the percentage of cuts at the dummy root for the different slots, model types and WSD options. A cut at the dummy root indicated that any preferences were not large enough for detection with the MDL technique. The percentage of (dummy) root cuts are displayed for the object slot in table 3.7, for the subject slot in 3.8 and for the PP slot in 3.9. In each of these tables the model types and WSD options are given.

FirstS reduced the number of root cuts observed for all slots and all model types. SPass did not do much to alleviate the problem of root cuts. This was because there were no preferences available for WSD from the first acquisition cycle for disambiguation of the argument head data. For ATCMs

Table 3.7: Percentage of root cuts with different WSD options, direct object

Model	NOWSD	SPass	FirstS	COMB
PTCMs	36	36 (08)	31	31 (16)
ATCMs	26	22	18	17
LLRTCMs	16	16	11	11

Table 3.8: Percentage of root cuts with different WSD options, subject

Model	NOWSD	SPass	FirstS	COMB
PTCMs	41	41 (13)	39	39 (15)
ATCMs	22	20	20	20
LLRTCMs	28	27	21	21

Table 3.9: Percentage of root cuts with different WSD options, PP

Model	NOWSD	SPass	FirstS	COMB
PTCMs	76	76 (21)	65	65 (40)
ATCMs	67	65	48	51
LLRTCMs	59	56	39	39

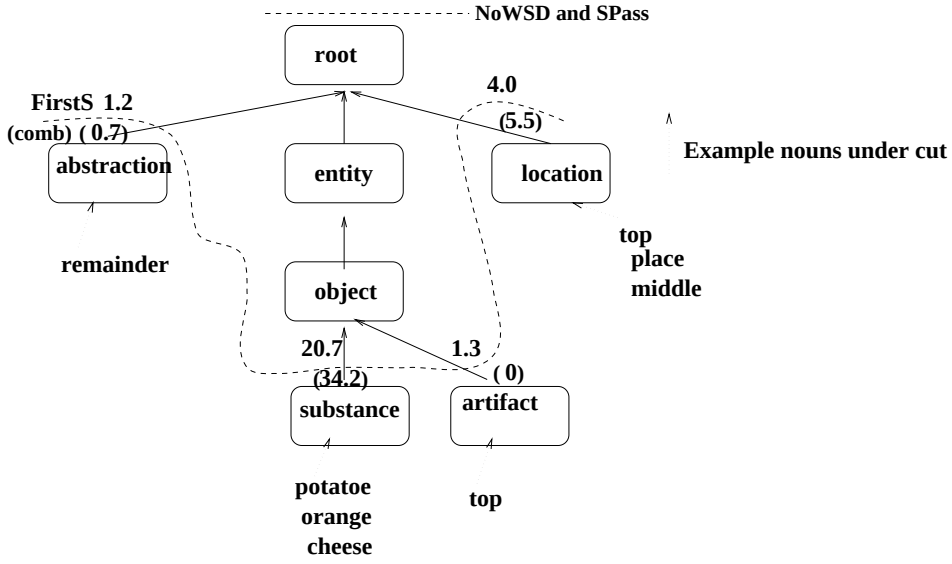

Figure 3.10: ATCM for *slice* object slot using the WSD strategies

and LLRTCMs, SPass did improve matters a little since the preferences that were obtained were applied to the data irrespective of the verb. This prior distribution was used in the preference scores and description lengths of both ATCMs and LLRTCMs.

Model Type

PTCMs were typically more general than LLRTCMs and ATCMs. There were many more root cuts for these models, regardless of slot. FirstS reduced the number of root cuts, however, a large proportion of verbs were left with root cuts even with FirstS disambiguation. For this reason, we experimented with a cut at the eleven root classes of WordNet (a 'WordNet Root Cut') in cases where MDL selected a cut at the (dummy) root above this. This is shown in figure 3.11, which illustrates the PTCMs obtained for the direct object slot of *scan*. The PTCM at A was obtained by the standard NOWSD setting. When this was used for SPass there was no improvement because there were no preferences to disambiguate the argument heads. When the WordNet root cut was used, we obtained a PTCM at B, with probabilities along this [7] from the conditional distribution $p(class|scan)$. When the PTCM at B was used for SPass, a PTCM at C was obtained. We obtained dramatic reductions in the number of root cuts output from SPass when we used the WordNet root cut for verbs which had a cut at the dummy root. We only experimented using the WordNet roots, instead of the dummy root, for the PTCMs. The percentage of dummy root cuts obtained using WordNet roots for the PTCMs are displayed in the tables in brackets. The verbs where a root cut was obtained selected less strongly for their arguments. It may be that dummy root cuts are better for NLP applications in these difficult cases by indicating where preferences are too weak to be reliable.

[7] The probabilities are not shown to the sake of clarity.

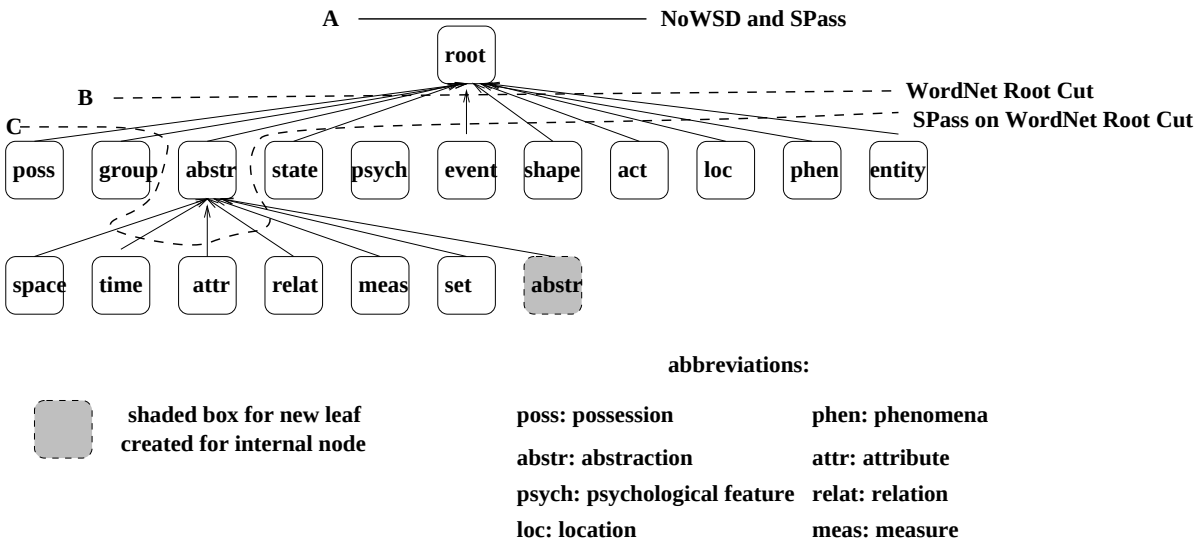

Figure 3.11: Using a WordNet root cut for *scan* object slot, for SPass WSD

Slot

The effect of WSD was the same regardless of slot. PPs were particularly prone to root cuts because of the sparse data problem. Data was collected not only with respect to the verb but also to the preposition. As Wagner (2000) has pointed out, the quantity of data is important when using MDL. If $|S|$ is the sample size then, given Li & Abe's formulations, the model description length has complexity O($\log|S|$) whilst the data description length has complexity O($|S|$). With increasing sample size the data description length increases more rapidly than the model description length. The model description length is more expensive in situations where less data is available. In these cases, there is a tendency for a more general model, and so cuts at the root are more common.

WSD helped considerably in cases of sparse data. Without WSD, 67% of our ATCMs at PPs were cut at the root. With FirstS this proportion was reduced to 48%.

3.6.3 Probability Distributions

The decrease in the percentage of root cuts demonstrates the increase in specificity provided by lexical disambiguation, at least for FirstS. We compared the actual probability distributions in WordNet to see the effect of the WSD options on the homogeneity of the data. We did this using the probability distribution at the root classes in WordNet for the direct objects of a couple of verbs (*produce* and *melt*) used as examples in section 3.6.1. The selectional properties of these verbs differed somewhat. *Melt* had a strong selection for classes under the **substance** class, whereas *produce* had more diverse preferences.

The frequency distributions are shown for all WSD options. SPass was performed using the ATCMs. [8]

[8] Note that the probability distribution across the roots do not sum to 1 because of errors in rounding, and any overlap of classes caused by multiple parentage.

Table 3.10: Probabilities at root classes - *melt* direct object

Root (synonyms at)	NOWSD	SPass	FirstS	COMB
possession	0.003	0	0	0
group grouping	0	0	0	0
abstraction	0.1	0.02	0.1	0.05
state	0.003	0	0	0
psychological_feature	0.04	0.01	0.03	0.01
event	0	0	0	0
location	0.02	0.06	0.01	0
shape form	0.01	0	0.01	0
act human_action	0.01	0	0	0
phenomenon	0.01	0	0.01	0
entity	0.8	0.92	0.8	0.94

SPass worked well with verbs that select strongly for their objects. For these verbs, ambiguities were resolved in favour of senses which fell in the predominant semantic area. For *melt*, SPass provided a larger increase in probability at the **entity** class than FirstS, although both produced increases compared to NOWSD. COMB produced the largest increase at this class. The probability distribution conditioned on *produce* also showed a probability increase in areas of preference after disambiguation with SPass. This can be seen at the **entity** and **abstraction** classes. SPass, by its very nature, always increased the concentration of frequency in areas of preference.

3.6.4 Description Lengths

In this subsection, the description lengths (or relative costs) of the TCMs using the different WSD options are compared. As a consequence of using MDL, for each TCM we had a final description length associated with the TCM. This was minimised in our search for the optimal model. The optimal model, according to MDL, is the one which makes the best compromise between being succinct, and reflecting the data well.

In a pure MDL approach, the cost of a model is the actual description length, measured in bits, of both the model and the data when encoded in the model. As described in the previous chapter, a PTCM description length is clearly related to the number of bits required since we use $\sum_{class \in cut} \log p(class)$ for the description length. The description length for an ATCM is rather convoluted as we use the log of the association score in the description length, having envisaged the ATCM as a by-product of a process to produce a TCM for the conditional distribution. The description length we use for a LLRTCM is not clearly related to the number of bits required for description since the LLR statistic is used as a heuristic, in place of a clear description length. The costs used in the acquisition process for ATCMs and LLRTCMs do not equate to the actual number of bits required for encoding the data in our model. They do, however, provide relative costs which are sufficient for the purpose of ranking models. Using the same description length calculations (i.e. the same model type) on the same data (i.e. for the same verb), we compared not only the

Table 3.11: Probabilities at root classes - *produce* direct object

Root (synonyms at)	NOWSD	SPass	FirstS	COMB
possession	0.02	0.01	0.02	0.001
group grouping	0.04	0.03	0.04	0.04
abstraction	0.3	0.4	0.2	0.3
state	0.04	0.04	0.04	0.03
psychological_feature	0.11	0.07	0.11	0.08
event	0.04	0.02	0.03	0.01
location	0.02	0.002	0.01	0.01
shape form	0.007	0.001	0.005	0.002
act human_action	0.1	0.03	0.1	0.1
phenomenon	0.05	0.03	0.09	0.1
entity	0.3	0.4	0.3	0.4

Table 3.12: Average cost

Model	NOWSD	SPass	FirstS	COMB
PTCMs	11172	10025	10782	10335
ATCMs	-953	-1638	-1162	-1442
LLRTCMs	-565	-1860	-830	-1365

resultant model, but its description length for the different WSD options. We only compared costs of models for the same verb, or across the same set of verbs. Table 3.12 contrasts the average of the final costs for the set of verbs in the sample. These are the verbs which were provided with preference models (non-root cuts) for all WSD options for the direct object slot.

The final average costs for the ATCMs and LLRTCMs were negative. This is because costs are minimised when preference scores are maximised. The association score and the log-likelihood ratio score are negated in the equations (the ATCM equation 2.20 is on page 34, and the LLRTCM equation 2.21 is on page 38).

For all model types the final cost was reduced by WSD. This indicated an increase in the homogeneity of the argument head data after WSD. SPass provided a larger decrease in cost because selectional preference disambiguation, by its very nature, placed more data in the regions with highest concentration.

This subsection, and the preceding two subsections, provide quantification of the increased frequency concentration when WSD was applied to the argument head data used for preference acquisition. The effect was reduced for verbs where the selectional preferences were less marked. We evaluate whether or not the increased concentration translates to improved performance of the preferences in the next chapter.

3.7 Conclusions

Acquisition of selectional preferences has been performed on ambiguous data, apart from some relatively small scale experiments (Ribas, 1995a; Pozanski & Sanfilippo, 1996). Ribas (1995a) reported the noise from erroneous senses as a significant problem for selectional preference acquisition, and a source of over-generalisation of the preferences. In this chapter, we looked into WSD systems with a view to disambiguating the argument head data. Many WSD algorithms have only been applied to a small sample of target words. Performance varies dramatically depending on the test data. It is not possible to predict performance on the all nouns task from results reported on a small test set. Furthermore, overheads, such as human supervision and machine training time, are important factors when applying WSD on a large scale.

From experimenting with three WSD options, we selected two for disambiguating the nominal argument head data. These two options are (i) SPass — using the selectional preferences acquired from the ambiguous data in a second pass cycle and (ii) FirstS — using the first sense of any word, regardless of context, where there is evidence that this sense is predominant. These techniques did not match the precision and recall figures cited in the WSD literature, however they did outperform the random baseline. They were easily applied to the nominal data without prohibitive overheads.

SPass increased the intensity of preferences, and decreased weak associations elsewhere. This was shown by comparing probability distributions at WordNet roots with and without disambiguation. Both WSD options reduced the description length costs of the TCMs, suggesting that the argument head data set was more homogeneous after disambiguation. This increase in the homogeneity of the probability distributions brought about an increased specificity in the preference models. The specificity of models was increased with both options. For the majority of cases, like *produce*, FirstS provided more specific models than SPass, because FirstS disambiguation was more precise. FirstS could operate in cases where SPass could not since some verbs had TCMs at the root. For this reason FirstS was better able to reduce the number of root cuts over the full set of verbs. We also used the two WSD options together for the COMB option. For this option, we used the preferences only where there was not a clear predominant sense for the target lemma. Formal evaluation of the WSD options is reserved for the next chapter.

Chapter 4

Evaluation of Automatically Acquired Preferences

4.1 Introduction

This chapter concerns the formal evaluation of automatically acquired selectional preferences. We evaluated our selectional preference models to see how they compared to those produced using alternative methods and to see how different parameter settings and model types (ATCM, PTCM and LLRTCM) affected performance.

The three model types were introduced in chapter 3. They result from differences in the preference measure, which give rise to different description length calculations. The description length is used in determining the correct level of generalisation in WordNet. The ATCMs use a measure based on mutual information, the PTCMs use conditional probabilities ($p(c|v)$) and the LLRTCMs use LLR. The parameter settings include the WSD options (SPass, FirstS and COMB) introduced in the last chapter. These options relate to disambiguation of the argument head data using:

1. SPass – the selectional preferences
2. FirstS – the first sense heuristic
3. COMB – a combination of both

In this chapter, we also discuss the results of one evaluation on TCMs which were obtained using three other strategies. In the first of these strategies, we used Li & Abe's (1996, 1995) original method of pruning WordNet at classes featuring lemmas which occurred in the argument head data. This contrasted with our strategy of creating new leaves for word senses at internal nodes. Secondly, we evaluated the effect of identifying, rather than ignoring, proper nouns. Thirdly we evaluated ATCMs obtained from the argument head data specific to a SCF and slot combination, rather than simply to a slot.

In chapter 2 we contrasted these model types and the effect of these strategies informally. We looked at the preference model characteristics in terms of their specificity and the proportion of preferences which were at the dummy root. In chapter 3, we made these comparisons for the WSD options. We also looked at the effect of WSD on the probability distributions and the

description lengths. Our conclusions were that WSD increases the homogeneity of the preferences, and therefore reduces the description lengths of the TCMs.

In this chapter, we use formal quantitative methods for evaluation. Research on the lexical acquisition of SCFs has made use of various gold standards (Manning, 1993; Briscoe & Carroll, 1997). There are several dictionaries available with SCF information (Hornby, 1989; Boguraev et al., 1987; Grishman et al., 1994). Evaluation of selectional preferences is difficult since there is no clear gold standard. Selectional preferences are semantic, and the appropriate categories for describing them are not obvious. Even given a taxonomy for classification, the semantic constraints on arguments are less obvious than syntactic ones. This is exacerbated by the fact that semantic constraints are only preferences, rather than hard and fast constraints. Words can be used in novel ways, both syntactically and semantically. Arguably, speakers conform to syntactic constraints rather more than they do to semantic ones. For these reasons, we would expect less accordance between the decisions of different lexicographers producing selectional preference entries, compared to SCF entries. It is difficult for a human to say precisely what the preferences for a verb should be (Fillmore, 1970), without making an explicit list of the lexical fillers. For verbs dealing with concrete nouns, for example *drink*, it may be relatively easy. However for verbs taking abstract nouns, *explain* for example, it is much harder to define preferences in more detail than by stating that they take an **abstract** category.

Selectional preferences are acquired for a wide variety of different purposes. Their representation and evaluation will reflect this. For example, a selectional preference acquisition system designed specifically for speech recognition will pay more attention to the conditional probability $p(word|verb,slot)$ than our system does, because distinctions need to be made at the word level. Selectional preferences designed for WSD will need to predict $p(sense|verb,slot)$. The preferences produced by our system were required for diathesis alternation identification. We used a class-based system for this, since generalisation to classes reduced the problem of sparse data. Systems with other applications in mind also make generalisations for this reason. We describe the evaluation of our preferences on the task of diathesis alternation identification in the next chapter. This chapter concerns other formal evaluations we have performed on our models. these evaluations were done to compare our preference models to those of other researchers, and to investigate how the various parameter settings affect performance.

In the next section, we provide a broad categorisation of evaluation techniques for lexical acquisition and see how these methods might be employed for selectional preference acquisition. In section 4.3 we look at evaluation methods for automatically acquired selectional preferences which have been reported in the literature. We do this in light of the categorisation provided in section 4.2. Section 4.4 describes the evaluations that we have performed on our models, and the results obtained. This is followed by the conclusion of our results in section 4.5.

4.2 Evaluation Methods For Lexical Acquisition

In evaluation of lexical acquisition systems, a distinction is sometimes made between type-based evaluation, and token-based evaluation (Briscoe & Carroll, 1997). Types are the entities being acquired, the entries made in the lexicon. These can be evaluated by comparing the types acquired against those provided in a gold standard. The gold standard is typically compiled by humans.

Tokens, meanwhile, refer to corpus instances which are manually analysed and compared to the acquired information.

Type-based evaluation requires a gold standard. For selectional preference evaluation, the gold standard might take the form of precompiled selectional preferences. As an alternative one might use a team of judges to decide whether the acquired preferences were adequate. The problem with using human judges is one would need to supply randomly produced preferences to provide some sort of a baseline. However, randomly produced preferences are unlikely to look reasonable because of the large space of possibilities with a large scale semantic taxonomy such as WordNet. The task would be too easy and there would be no easy way of deciding objectively where improvements could be made. To make the task harder, one might make the 'red herrings' more plausible, for example, by modifying automatically produced preferences for similar verbs. However, it would not be easy to define how far to make these false preferences look realistic. One further problem with type-based evaluation is that entries in a gold standard cannot be acquired if they are not attested in the corpus. Furthermore, corpus-based acquisition will acquire useful information which may have been omitted from the dictionary.

Token-based evaluation is usually performed on corpus data. It can be performed on the corpus data from which the acquired entities were obtained, to quantify coverage of the training data (Ribas, 1995a; Briscoe & Carroll, 1997). It might also be performed on a different corpus to see how well the acquired information generalises to a new data set. Analysis of the training data avoids the problem prevalent with type-based evaluation using a precompiled inventory, that the system cannot acquire information unless it is attested in the corpus. If unseen corpus data is used for testing, the new data may contain entities not attested in the training corpus, however, something can then be said about generalisation. For token-based evaluation of acquired selectional preferences, one might look at the predicate and argument head instances in the test data and see if they fall under the acquired selectional preferences.

The above evaluation strategies concern the 'objective' of the lexical acquisition system (Sparck Jones & Galliers, 1996). They show if the acquired information was that sought, and whether erroneous information was excluded. As such, these strategies relate to *intrinsic criteria* in the terminology of Sparck Jones & Galliers (1996). If we were interested in outputting a lexicon for lexicographic purposes, we would then ensure that the output was appropriate for human consumption. This is the 'function' intended for our system, referred to as the *extrinsic criteria* by Sparck Jones & Galliers. For NLP, we want to know that the information that we have acquired is useful. We therefore investigate the system's performance on relevant tasks. This is usually referred to as task-based evaluation.

Ultimately, task-based evaluation should be done on the fully fledged NLP system that the lexical information is required for. One way to evaluate the contribution of the lexical information is to compare the final system with and without this component, to see how much removing the component degrades performance (Gaizauskas & Humphreys, 1996). For example, selectional preferences might be used in an information extraction system which finds specific pieces of information from a text. The preferences might be used to help resolve anaphora, or perhaps to disambiguate words. The system would then be evaluated with and without the selectional preference component to see the benefits. This thesis concerns how acquisition of SCFs and selectional

preferences can be used for identification of diathesis alternations. The evaluation of our preferences on this task will be done in the following chapter. To allow some comparison with the preferences produced by others, we evaluate our preferences using some of the tasks that others have used.

4.3 Evaluation of Automatically Acquired Selectional Preferences: Previous Work

In this section, we discuss the evaluations reported in the literature that have been performed on automatically acquired selectional preferences. We start by commenting on the lack of type-based evaluations. We then describe two token-based evaluation strategies that have been devised by Ribas (1995a). Finally, we discuss the variety of task-based evaluations for selectional preferences.

4.3.1 Type-Based Evaluation

Aside from introspections provided by the researcher (Ribas, 1995b; Li & Abe, 1998; Resnik, 1992), there have been no attempts to use a gold standard for evaluating selectional preference entries. This is understandable, given the difficulty of obtaining a gold standard. There are selectional restrictions provided in the on-line version 1 of LDOCE. These have not previously been used for evaluation of automatically acquired selectional preferences. However, the selectional restrictions have been used by NLP systems (Wilks & Stevenson, 1998b). We have evaluated our selectional preference entries against the entries in LDOCE. We elaborate more on this in section 4.4.1.

4.3.2 Token-based Evaluation

Ribas (1995a) used a token-based approach to quantify the appropriateness of the generalisation level of his selectional restrictions. These were represented as classes in the WordNet hyponym hierarchy. We provided a description of his acquisition method on page 23 in chapter 2. Ribas devised a measure which he referred to as the 'generalization ratio'. [1] This required sense tagged material. It used the word senses for nominal argument heads in the specified slot in the corpus. The measure was calculated over all the selectional restrictions (predicate, slot and WordNet class combinations) where at least one word sense from the WordNet class was observed in the specified slot of the predicate in the test corpus. The generalization ratio was the number of senses that fell at or under this subset of acquired selectional restrictions divided by the total number of senses (incorrect or correct) that fell at or under the subset of selectional restrictions. In other words, the denominator summed over the possible senses of the nouns under the restrictions, disregarding the sense tags. To take a simple example, suppose the test corpus contained two tuples, <*eat, direct object, chicken*> and <*bake, direct object, cake*> and the only selectional restrictions applicable were SR(eat,direct object) = **food** and SR(bake,direct object) = **object**. *Chicken* only has one sense under **food** in WordNet, and this was the intended sense of the token <*eat, direct object, chicken*>. Its other senses do not fall under these selectional restriction classes. *Cake*, meanwhile, has three senses under **object**. These are i) the **baked goods** sense, (ii) the **patty** sense and (iii) the **bar of soap** sense. The generalization ratio would be $\frac{1+1}{1+3} = \frac{2}{4} = 0.5$. The numerator is the number of senses under the selectional restrictions, 1 for *chicken* in its **meat** sense and 1 for *cake* in its **baked**

[1] This is referred to as the 'abstraction ratio' in (Ribas, 1995b).

goods sense. The denominator is the number of possible senses of these word forms which fall under the selectional restrictions. One sense of *chicken* falls under **food** and all 3 senses of *cake* do. This measure is not very intuitive. It does not tell us anything more than a WSD task would. It also says nothing about the selectional restrictions which are not applicable because they do not cover any senses in the test corpus. Nor does it say anything about senses which do not fall under the acquired selectional restrictions.

Additionally, Ribas used <verb, slot, noun–sense> tokens collected from the training corpus to estimate the coverage of his selectional restrictions. Manually disambiguated argument head data was used to calculate a measure referred to as 'strong coverage'. This measure was defined as the proportion of <verb, slot, noun–sense> tuples in the data where the noun sense was subsumed by a selectional restriction, with the given verb and slot, divided by the total number of tuples. Ribas also defined a 'weak coverage' measure. This calculated the proportion of tuples in the data where *any* sense of the noun was subsumed by a selectional restriction, divided by the sum of all tuples. For example, if the corpus contained one instance of $<$ *bring*,*direct object*,*evidence* - **legal statement**> and the only relevant selectional restriction was SR(bring,direct object) = **information**, the selectional restriction would have covered the wrong sense of *evidence*. strong coverage would be $\frac{0}{1} = 0$, whilst weak coverage would be $\frac{1}{1} = 1$. Strong coverage gives an idea of how well the selectional restrictions cover the data. Weak coverage is not a useful measure as there is no requirement for selectional restrictions to cover the wrong senses of the data. The major benefit that weak coverage brings is that there is no requirement for sense tagged data. These coverage measures do not take into account inappropriate selectional restrictions. There is no penalty for a system which simply outputs selectional restrictions for all the WordNet root classes. The generalization ratio would be adversely affected by such a system, so this should be considered alongside strong and weak coverage.

4.3.3 Task-Based Evaluation

Researchers have favoured the use of task-based methods for evaluating preferences. Typically, preferences have been acquired for a specific task, and so it makes sense to evaluate them on that task. The tasks that have been predominantly used for evaluation of automatically acquired selectional preferences are WSD, structural disambiguation, and a decision task where the preferences are required to differentiate between genuinely co-occurring, and artificially combined word pairs. The latter method is referred to as pseudo-disambiguation by Rooth et al. (1999), and we will adopt this terminology. Rooth et al. (1999) also evaluated their distributionally-based classification on the task of smoothing, that is, providing probability estimates for new data.

Word Sense Disambiguation Evaluation

As we pointed out at the start of chapter 3, selectional preferences can be used for WSD. Many researchers have applied automatically acquired selectional preferences to the WSD task (Ribas, 1995b; Resnik, 1997; Federici et al., 1999). We have already discussed some of these results in our comparison of WSD systems in chapter 3. For completeness we include them again here. In chapter 3, we discussed possibilities for disambiguating the argument head data fed to the selectional preference acquisition system. In this chapter, we focus on the evaluation of the acquired selectional preferences.

Both Ribas and Resnik performed WSD on the sense tagged data in SemCor. Resnik (1997) tested selectional preferences acquired from the portion of the Brown corpus within the Penn Treebank. He used his preference models, described on page 27 in chapter 2 for the 100 most strongly selecting verbs found in the corpus data. These can be expected to perform better than verbs with weaker selectional properties. The senses of argument heads under the WordNet classes with the highest selectional association score were selected. Resnik achieved overall accuracy of 44%, averaged over the slot relationships, with a random baseline of 29%.

Abney & Light (1999) used the same training and testing set as Resnik, but with the HMM models described on page 25 in chapter 2. They obtained an accuracy for the direct object relationship of 42%, compared to Resnik's 44% for this slot. Using their automatic parser they obtained preferences for the 100 verbs acquired using the entire BNC as training data. For this set, accuracy increased to 54%.

Ribas obtained an accuracy figure of 53% using a subset of the SemCor data. This was not significantly better than the random baseline of 52%. This poor performance was undoubtedly affected by the small quantity of training data used. Since Ribas was comparing performance with and without sense tagging, he only used a portion of the SemCor data for training. The rest of the SemCor data was held out as test data. The training portion in both cases amounted to 20,000 tuples. In our experiments with lexicon A, we have 318,000 tuples for the object slot alone. Ribas obtained 56% accuracy for the preferences acquired from the same argument head data with sense tags included. This was an improvement on the result without disambiguation, but still only a little above the random chance baseline. The results of (Abney & Light, 1999) and (Ribas, 1995a) together imply that the quantity of training data used affects performance.

Ribas also evaluated on 2,658 manually analysed tuples which involved the verbs *rise, report, seek* and *present*. For this precision was 80% and recall 78%, compared to a random baseline of 63%. It should be noted that on this occasion the testing set used was a subset of the training set. [2] Ribas included monosemous nouns in his sample, thus giving rise to a higher baseline.

It is difficult to compare WSD performance in the literature since the test and training samples typically differ. The SENSEVAL (Rosenzweig, 1998) competition provided an opportunity for participants to compare system performance under the same conditions. Although, system differences will mean that those conditions may favour one type of system more than another. There were two systems using only automatically acquired preferences in the SENSEVAL competition. One of these was the SUSSEX system (Carroll & McCarthy, 2000). This used our ATCM preferences with FirstS WSD. For the all nouns task fine grained precision, to the HECTOR word senses, was 41%. The other system was the OTTAWA system (Kilgarriff & Rosenzweig, 2000). This system obtained a precision of 33% on the same all nouns task. The random baseline on this task was 30% with a phrase filter to handle the easy multi-word cases (14.6% without).

Federici et al. (2000) disambiguated Italian verbs using SCFs and selectional preferences acquired in the analogy-based system described in this thesis both on page 18 in chapter 2, and 59 in chapter 3. They used the equivalent of the SENSEVAL test suite for Italian, ROMANSEVAL. The example-base was built from MRDs containing subcategorization information. A portion of the verbs were sense-tagged, providing supervised training data. A precision of 81% was reported.

[2] He used the parsed version of the WSJ in the Penn Treebank.

This cannot be compared with the SENSEVAL results because of substantial differences in both the task and the data. Manning & Schütze (1999) point out that WSD of verbs is best done using subjects and objects, whilst nouns require a wider context.

Structural Disambiguation Evaluation

Many researchers have applied selectional preferences to structural disambiguation (Li & Abe, 1998; Resnik, 1993b, 1993a). This has typically been performed to resolve PP attachment using the parses in the WSJ section of the Penn Treebank as training and evaluation data. The task involves resolving the ambiguity in sentences like the one shown in example 12. The selectional preferences are used to determine whether PPs, (*with the stick* in our example) should be attached to the verb (*hit*) or the NP (*the man*).

(12) The boy hit the man with the stick

The parsed data is used to collect tuples where the PP is attached to the verb, and those where the attachment is to the NP. Selectional preferences are obtained for the head nouns in the two types of PP. The preferences for the verb attached PPs are specific to the verb, *hit* in our example. The preferences for the NP attached PPs are specific to the head of the NP, *man* in our example.

The test data is of the form <verb, noun1, prep, noun2>. The preferences are used to obtain two scores: (i) a score for the preference for noun2 given <verb, prep >, using the verb attachment preferences, and (ii) a score for the preference for noun2 given <noun1, prep >, using the noun attachment preferences. These two scores are then compared.

Resnik (1993b, 1993a) used mutual information scores collected over WordNet for both the verb and NP attached data. There was ambiguity in the data because the nouns can be classified under many different WordNet classes. For the NP attached data, Resnik took each value of noun2 in turn and found the value of noun1 that maximised the mutual information score, obtained using his method of populating WordNet with probability distributions, which we described earlier on page 22 in chapter 2. The classification of noun1 (class1) was then fixed for the value of noun2. The mutual information scores for all possible classes of noun2 ($class2 \in CLASSES_{noun2}$) for the NP attached data were multiplied by the frequency of <class1, prep, class2>. The mutual information scores for the verb attached data were obtained in the same manner, except that noun1 is not involved, the verb is fixed, and the mutual information scores over the classes of noun2 are multiplied by the frequency of <verb, prep, class2>. A paired samples t-test was used on the means of the mutual information scores over the possible values of class2, from the NP attached, and verb attached data respectively. If the value of t was positive then the target was resolved in favour of NP attachment, if the value was negative VP attachment was used for the target instance. Resnik (Resnik, 1993a) obtained an accuracy of 79% using this method, when his system was left to decide in all cases. Accuracy was a little lower than Hindle & Rooth's (1993) method using lexical information. However, Resnik showed that if confidence levels were used to restrict application, his class-based approach had wider coverage.

Abe & Li (1996) applied their tree cut models to PP disambiguation. They also used the attachment decisions of the parsed portion of the WSJ in the Penn Treebank. They obtained ATCMs using the process which we described on page 31 in chapter 2. They then compared the association scores for the possible classes of noun2 on the ATCM for verb attachment, with the scores on the

ATCM for NP attachment. They also obtained results on the same data for Resnik's selectional preference models (Resnik, 1993b) and their conditional probability models (akin to our PTCM models). Precision was around 95% for all methods, but the methods achieved different levels of coverage. The ATCMs provided 80% coverage whilst the selectional association models covered only 64% of the data. The probabilistic tree cut models achieved 73% coverage.

4.3.4 Pseudo-Disambiguation Evaluation

In this task, the system has to distinguish which of two word pairs is a more likely co-occurrence. The task is typically performed with respect to specific slot relations for a verbal predicate. The tuples comprise the slot, verb, and lemma in argument head position at the slot. A test set of correct tuples are extracted from a corpus. Each tuple is given to the system alongside an artificially produced tuple. These artificial tuples are often produced by replacing the verb lemma from the genuinely occurring tuple, with a verb selected at random. The exact mechanism of producing the artificial pairs varies, and this can affect results.

Tuple evaluation has been favoured by researchers using proximity based methods (Grishman & Sterling, 1993; Pereira et al., 1993; Lee, 1999; Rooth et al., 1999). Rooth et al. (1999) used a tuple decision task to evaluate their automatically produced classifications. The semantic classes in the classifications contained both verbal predicates and noun arguments and were described in chapter 2 on page 17. The tuples comprised two verbs (v and v') and a noun (n) which were all seen in isolation in the training corpus. Only one of the verbs (v) was attested in the test corpus with the noun. Additionally, both vn and $v'n$ combinations were removed from the training corpus. The invalid verbs were selected with regard to frequency. All the lemmas used in evaluation occurred between 30 and 3000 times in the training corpus, in a specified slot relationship. There were 1337 evaluation triples and accuracy was calculated as the number of times that the probability estimate for the genuine pair (vn) was greater than that for the unattested pair ($v'n$). The random baseline was 50%. Optimal performance at 80% accuracy was obtained for models with between 25 and 100 classes.

(Pereira et al., 1993) performed a similar evaluation on their distributional classification. This was conducted using only 104 verb-noun pairs selected at random from the 44 million word 1988 Associated Press newswire corpus. Only verbs with a frequency between 500 and 5000 were chosen. The test pairs (vn and $v'n$) were removed from the training corpus. The system's decisions, on which pair was more likely, were compared to those provided by the frequencies of these test pairs in the original training corpus. The proportion of errors was around 23% (77% accuracy) when the model included more than 50 classes.

Grishman & Sterling (1993) evaluated their selectional constraints on a slightly different tuple task. The selectional constraints in this case were confusion matrices obtained from distributional data. We described these in chapter 2 on page 18. Instead of using artificially created tuples for evaluation, they obtained tuples from all the possible parses of the test corpus using a non-stochastic grammar. The tuples were manually classified as valid or invalid for evaluation purposes and the systems decisions were compared to the decisions of the human experts. The tuples were scored as true positives (TPs) if the system and the experts both recorded the item as valid. False positives (FPs) occurred if the system alone marked the item as valid. The items were scored as

true negatives (TNs) if the system and experts both agreed that the tuple was invalid, and as false negatives (FNs) if the item was marked as invalid by the system, but as valid by the experts. Recall was calculated using $\frac{TP}{TP+FN}$. The error rate was calculated as $\frac{FP}{FP+TN}$. A 'quality measure', given in equation 4.1, was used to see if the smoothing process performed better than a random process would have been expected to. In the equation, the *s* subscript represent the values after smoothing and the *ns* subscript represent the values using the raw training data for predicting validity, without any smoothing.

$$\text{Quality measure} = \frac{\frac{TP_s - TP_{ns}}{FN_{ns}}}{\frac{FP_s - FP_{ns}}{TN_{ns}}} \tag{4.1}$$

Performance of the confusion matrices was also compared to that using a manually created hierarchy, produced for the Fourth Message Understanding Conference (MUC-4) (MUC-4, 1992). According to the quality measure, both the confusion matrices and the manually produced hierarchy produced better results than those predicted for a random process for smoothing. Smoothing with both the manually created hierarchy and the confusion matrices improved the recall, at the expense of error rate, compared to the non-smoothed model. The confusion matrices obtained a recall of 34% and an error rate of 9% when using a threshold (0.29) on the frequencies provided by the confusion matrices produced. The non-smoothed model achieved 30% recall and 7% error rate. The results for the confusion matrices were not significantly better than those for the manually produced hierarchy.

4.3.5 Smoothing

Rooth et al. (1999) also used a measure of 'smoothing power' for their distributional-based classification models. This concerned the coverage of test data with the acquired classification. Rooth et al. calculated the proportion of 1000 randomly selected verb and noun pairs which were assigned a joint probability by their models. Coverage of this test data was increased by reducing the number of classes allowed, with the highest smoothing power associated with a classification containing only 1 class. A compromise was required between performance on this task and discriminatory performance of the classes on the pseudo-disambiguation task. A model with 50 classes achieved a good compromise and attained a smoothing score of 93%.

4.4 Evaluation of the TCMs

In section 4.2, we outlined three approaches to evaluation, (i) type-based (ii) token-based and (iii) task-based. In section 4.3, we provided examples of each from the literature on automatic acquisition of selectional preferences. In this section, we describe the evaluation of our TCMs using all three of these approaches. The purpose of our evaluation was threefold. Firstly, to see how our models compared with semantic constraints specified a priori by lexicographers. Secondly, to see how the performance of our models compared to those of other preference models reported in the literature. And thirdly, to see how the parameter settings affected performance.

For type-based evaluation, we compared our preference models to preferences provided in the on-line version of LDOCE. This was performed using a manually created link between the

Table 4.1: The percentage of argument head lemmas not in WordNet

slot	% not in WordNet
object	3
subject	5
PP	4

WordNet hierarchy and the LDOCE semantic codes. Token-based evaluation was performed using the tokens from a dictionary, the Cambridge Dictionary of International English (CIDE) (Procter, 1995), rather than a corpus. For the CIDE evaluation, we manually tagged the argument heads in a small set of dictionary examples with WordNet senses. We then calculated the proportion of these examples that were associated with preferences by our TCMs.

Using dictionaries for evaluation does have its problems. For both type-based and token-based approaches, there are likely to be significant differences between the facts that the lexicographer feels are important, and naturally occurring corpus data. Both the entries in LDOCE, and the dictionary examples in CIDE are likely to contain a higher proportion of specialised senses and rare usages. Many of these will not be attested in corpus data used for obtaining TCMs. TCMs will be penalised in these cases. Additionally, there are likely to be senses in corpus data that are absent from the dictionaries. Nevertheless, we performed the dictionary-based evaluations to see the extent of coverage of the lexicographers' examples, and to observe the behaviour of the various parameter settings and model options with regard to this.

An alternative approach for token-based evaluation is to investigate the coverage of manually analysed data. Ribas (1995a) did this using a manual analysis of a portion of the training data. He used this for his generalization ratio, and strong coverage measures described in section 4.3.2 above. The generalization ratio calculated the proportion of correct senses from the data that were under classes with a preference, compared to the proportion of both valid and invalid senses for the subset of sense tokens which occurred under an acquired selectional restriction. The strong coverage measure calculated the proportion of sense tokens which occurred under a selectional restriction. The generalization ratio is particularly unintuitive, and we feel it is less informative than performance on a WSD task. It does, however, favour discriminatory preferences. The strong coverage measure was used on the training data and the score obtained simply reflected the coverage of this. It would have been interesting to know how this related to coverage of held-out data.

Instead of manually tagging a portion of the training data with WordNet senses, which would have been a laborious process, we relied on recall in the two task based evaluations, and the number of verbs with root cuts to provide an insight into coverage of the *test* data. In addition to this, the percentage of lemmas from lexicon A not found in WordNet for the subject, object and PP slots is given in table 4.1. The two task-based evaluations that we performed were WSD and the pseudo-disambiguation task.

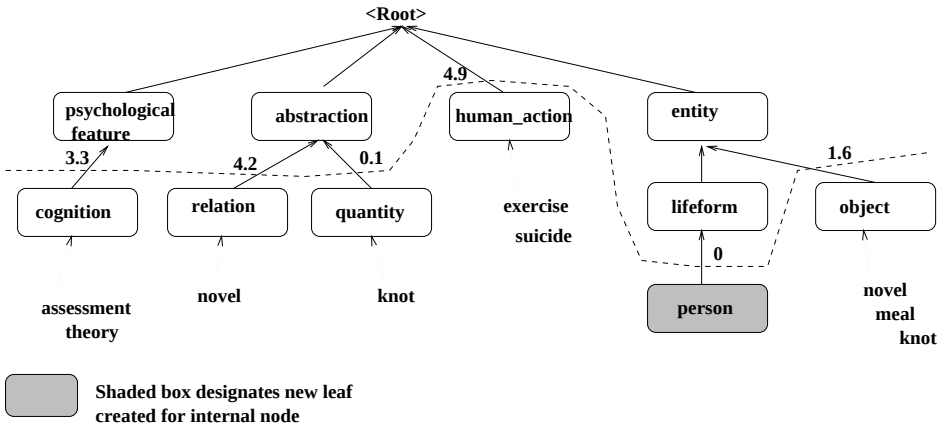

Figure 4.1: ATCM for *attempt* direct object slot

4.4.1 LDOCE Evaluation

The MRD LDOCE (Procter, 1978) version 1 includes semantic codes which provide constraints on the semantic type of fillers for the subject and object slots. These constraints are expressed as hard and fast restrictions, rather than preferences on a continuum. There are 32 semantic codes which are made up of combinations of 16 core categories. One of the 32 codes denotes **no restrictions**. The labelling of the subject and object slots for the verbal entries was provided by human lexicographers. Only one semantic code is specified for each slot and verb sense combination. The semantic codes are broad and cover a multitude of slots fillers. For example, the class **W** is used for a non-animate category, it covers both inanimate and abstract items.

The 'gold standard' falls prey to errors, inconsistencies and omissions as one would expect from such a large manmade resource. For example, the category of **no restrictions** is placed on the direct object of the first sense of *attempt*, the sense of *to make an effort at*. The glossary provided within this version of LDOCE is as follows:

> to make an effort at; try : He attempted the examination but failed. I attempted to speak but was told to be quiet. I attempted walking until I fell over. He was found guilty of attempted murder even though the other man did not die.

It would seem that a preference for the **abstract** LDOCE code would be more appropriate. This would express a difference between the acceptability of *attempt an examination* from less likely combinations such as *attempt a woman*. A portion of the ATCM (without any WSD) obtained for *attempt* is illustrated in figure 4.1. The areas of preference (with scores above 1) clearly accord with intuition, and with the examples in the LDOCE glossary. In many cases where the category of **no restrictions** is applied, it may be that this is done because the broad nature of the LDOCE semantic codes make it difficult to make subtle distinctions. For example, *draw* (the **with a pencil** sense) and *sing* (the **song** sense) both have the **abstract** class for the direct object slot. Whether this is a problem really depends on the frequency with which the restrictions are required to make narrower distinctions. The LDOCE restrictions have been used successfully for WSD. Wilks & Stevenson (1998b) made use of them in their WSD system. They obtained an accuracy of 57% when using these alone on a small portion of the SemCor data (200 words). This sample included

monosemous words and was compared to a baseline of 50%, if the first sense of each word was selected for the training and test sample put together.

On the assumption that these manually produced restrictions have some validity, we compared our automatically produced preferences to see (i) the extent to which the acquired ones were matched in LDOCE and (ii) the extent to which the LDOCE preferences were found by our system. We did so by calculating precision and recall over the types contained in LDOCE. These measures are defined in equations 3.2 and 3.3 on page 52 in terms of true positives (TPs), false negatives (FNs) and false positives (FPs). TPs were the cases where a restriction listed for a verb and slot in LDOCE also occurred in the TCM models produced by our system. FNs were cases where an LDOCE restriction did not occur in our model. Finally, FPs occurred when an acquired preference was not listed in LDOCE. To overcome the fact that our models contained preferences on a continuum signified by the preference scores, rather than restrictions, we used a threshold on the preference score. Only preferences with a score above the threshold were compared to the LDOCE restrictions.

Of course, the dictionary and corpus-based preferences were expected to differ, regardless of the manmade versus automatic distinction. Corpus-based methods only observe preferences which appear in the training data. Moreover, using statistical techniques, as we do, we would only expect to find preferences that are reasonably common in the corpus data. LDOCE provides evidence for all senses of a verb, however rare. This lack of coverage of rare events by the TCMs was expected to adversely affect recall in this type-based experiment. This should be borne in mind when interpreting the results. It is better that the preferences cope with more common events, since the preferences were ultimately created for handling naturally occurring corpus text. A related issue is that acquisition of preferences from real data was expected to uncover legitimate preferences which have been omitted from LDOCE, either from error, or because the corpus contained specialised senses of the verb usages. This was expected to affect precision, and again was explained by the difference between dictionaries and corpora.

In order to compare the LDOCE restrictions with the classes in our TCMs, we required a mapping between LDOCE and WordNet. Although there is at least one such mapping in existence (Knight & Luk, 1994), this was not a resource available to us. We developed our own mapping for this purpose. We shall refer to this hereafter as the LDOCE-WordNet mapping. In our mapping, the LDOCE semantic codes were mapped to appropriate WordNet classes. This was a 'one to many' mapping between the core 16 LDOCE categories and WordNet, since invariably the LDOCE codes corresponded to several WordNet classes. Some categories, for example **animate**, **inanimate** and **human**, were easily mapped to WordNet classes. Others, for example **immovable solids** and **movable solids**, were not easily identified with a small set of WordNet classes. The two classifications make different distinctions. For some problematic LDOCE categories, such as the **immovable** - **movable solids** distinction, we simply mapped to a more general WordNet class which covered both categories, even though some distinctions would be lost. This was preferable to enumerating ungainly sets of WordNet classes. Some of the LDOCE categories refer to disjunctions of the core set of 16 classes. These were easily dealt with by mapping between the LDOCE category and all the relevant WordNet classes. For example, the category **animal or human** is mapped to the WordNet classes **animal** and **person**. Other LDOCE categories are defined using a conjunction of core categories, for example a category for **abstract and solid** is provided.

Table 4.2: LDOCE-WordNet mapping for some imaginary verb sense entries

LDOCE			WordNet
verb	verb sense	restrictions	WordNet classes
v1	$v1_1$	abstract	abstract event shape ...
v1	$v1_2$	abstract&solid	NIL
v2	$v2_1$	inanimate	object
v2	$v2_2$	movable	object
v2	$v2_3$	human	person
v3	$v3_1$	animal or human	animal human
v3	$v3_2$	human	human

Example verb entries with this restriction include:

- the direct object of *improve*, in the sense of **improve ones ability**
- the direct object of *conduct* in the sense of **direct the course of an activity**
- the direct object of *hold*, in the sense of **a container holding a quantity of a substance**

It is hard to see how the potential slot fillers for this category are related, and how they are distinguished from nouns with the general LDOCE **abstract** category. We did not evaluate on verbs for which the LDOCE restrictions include highly problematic LDOCE categories for which a reasonable mapping could not be found. The majority of category labels were *not* problematic. Furthermore, most of the verbs in our sample did not contain the problematic categories (86% of verbs having the direct object slot, and 90% for verbs having the subject slot).

Figure 4.2 displays a portion of our LDOCE-WordNet mapping. In table 4.2, we provide some imaginary examples of verb sense entries, with LDOCE semantic codes and the corresponding WordNet classes to show how our mapping worked. Note that the verb *v*1 would have been ignored in our evaluation because sense $v1_2$ includes a problematic category which cannot be mapped. Since LDOCE provides entries for verb senses, rather than forms, we included duplicated restrictions for a form as many times as they occurred for the senses of that form. Both *v*2 and *v*3 from our contrived example have duplicated restrictions.

We did not expect the TCMs to be at exactly the same level of WordNet as that specified by the LDOCE entries and the LDOCE-WordNet mapping. Differences between the LDOCE and WordNet semantic taxonomies made it hard to predict whether the TCMs would be above or below the LDOCE restrictions, when the latter were mapped to WordNet classes. On the whole, LDOCE preferences were expected to be higher than those in WordNet because there are only 32 LDOCE semantic codes, whereas there are in excess of 60,000 WordNet classes in the noun hyponym hierarchy of WordNet version 1.5. There are, however, counter examples. Some LDOCE semantic categories were mapped to rather specific WordNet classes, further down the hierarchy than the typical level of the TCMs. For example, the category **organic material** was mapped to a subset of the hyponyms of the WordNet **substance** class. We categorized LDOCE restrictions as TPs when the mapped class was either above or below a class on the TCM with a preference score above the

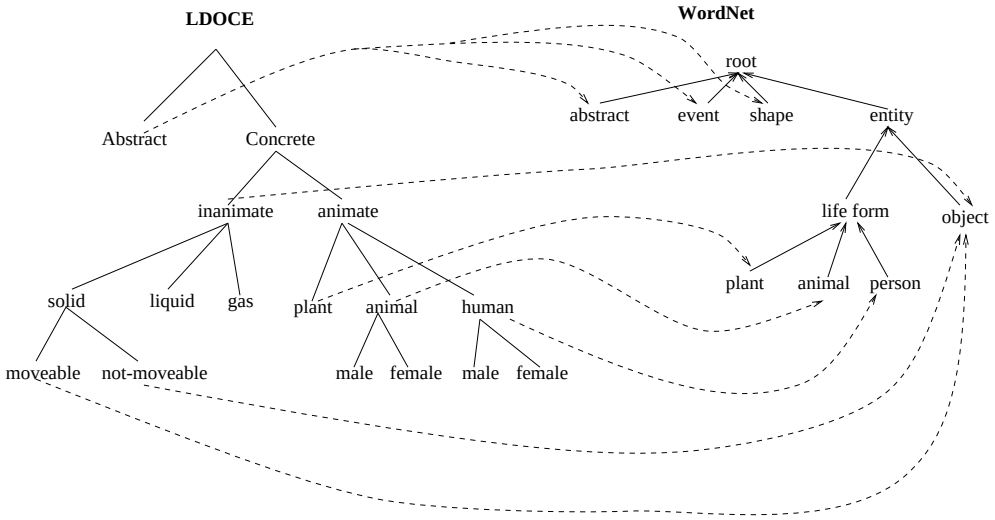

Figure 4.2: An illustration of the mapping between LDOCE and WordNet

threshold. The LDOCE restriction was counted as a FN otherwise. The acquired preferences, above the threshold, which were not above or below the mapped LDOCE restrictions were categorized as FPs.

There was a one-to-many mapping between LDOCE restrictions and the WordNet classes. The TPs and FNs were counted with respect to LDOCE restrictions. We needed to count FPs using the same LDOCE categories which were used for calculating TPs in order to combine the two counts for precision. The acquired WordNet preferences for a verb were combined together to map to LDOCE semantic codes in order to count the TPs. This should have ideally been done by searching the hierarchy for the smallest set of LDOCE categories to fit a given set of WordNet classes (which might be anywhere in the hierarchy). Instead, to avoid heavy computation, we used some heuristics to perform the mapping for the most prevalent cases.

We evaluated on the acquired TCMs for a set of verbs from a randomly selected sample of 500 sentences, intially referred to on page 36. The verbs were acquired using the data in lexicon A. The bulk of our evaluation was performed using the LDOCE restrictions at the object slot, we also compared performance for the subject slot using one parameter setting. Table 4.3 displays the precision and recall results for various TCMs. The TCM type is specified in the first column. The threshold (thresh) used depended on the TCM type. For ATCMs, an association score of more than 1 occurred when the conditional probability $p(c|v)$ was greater than the prior probability $p(c)$. This indicated a positive association between the verb and the class and therefore provided an obvious choice of threshold. However, a higher threshold did increase the precision, whilst reducing recall. For LLRTCMs a threshold of 0 was used. This was the point where the observed data for the conditional distribution was greater than that expected. We also tried a threshold of 1, and one of 3.84. The latter is the threshold for 95% significance (one tailed) when using the log-likelihood ratio against the chi-squared tables. The higher thresholds did increase precision, however, recall was still lower than that for the ATCM with a threshold of 1. For the PTCMs there was no obvious choice of threshold. Using a threshold at 0.1, we obtained similar results to those obtained with the ATCM and a threshold of 2.

Table 4.3: LDOCE evaluation: for different model types and thresholds

TCM	thresh	WSD	slot	precision	recall
ATCM	1	none	obj	63	75
ATCM	2	none	obj	70	64
LLRTCM	0	none	obj	45	71
LLRTCM	1	none	obj	56	70
LLRTCM	3.84	none	obj	63	67
PTCM	0.3	none	obj	89	45
PTCM	0.1	none	obj	70	64
ATCM	1	none	subj	60	69

Table 4.4: LDOCE evaluation: for ATCMs with different WSD options

TCM	thresh	WSD	precision	recall
ATCM	1	none	63	75
ATCM	1	SPass	65	73
ATCM	1	FirstS	57	80
ATCM	1	COMB	56	77

Using the ATCMs for the direct object slot, we evaluated the three WSD options introduced in the last chapter. The results are displayed in table 4.4. SPass slightly increased precision whilst reducing recall, however these differences were not significant. FirstS significantly reduced precision and increased recall.

Overall there was a reasonable level of coverage of the semantic constraints in the 'gold standard' (up to 77% recall). There were a considerable proportion of FNs caused by TCMs cut at the root. For example, 79% of FNs for the ATCM with a threshold of 1 and no WSD, were caused by TCM root cuts. SPass and FirstS reduced the number of these root cuts, as seen by the second column in table 4.5. Some LDOCE restrictions were not observed in the acquired preferences because they referred to specialised senses, which are less likely to occur in the corpus data. For example, the direct object slot of *continue* is given a **human** restriction for the sense specified with the glossary:

> to (cause to) stay in a particular job or office : The king decided to continue Pitt as chief minister

Precision values were adversely affected where acquired preferences were found which could not be mapped to the LDOCE ones. Some of these occurred where LDOCE provides the code for **no restrictions**. The percentage of verbs where this occurs is displayed in the third column of table 4.5. In many cases, the acquired preferences were appropriate. For example, a portion of the ATCM obtained for the direct object slot of *begin* is displayed in figure 4.3. This ATCM was acquired with no WSD options. The preference at the classes **action** and **activity** contrasts with

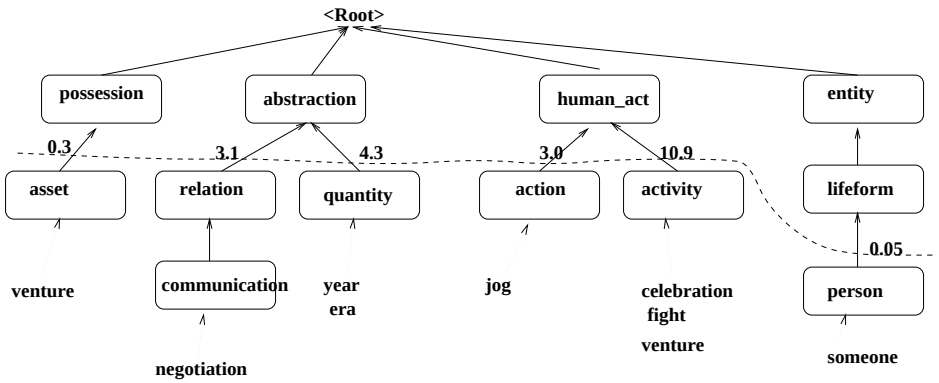

Figure 4.3: ATCM for *begin* direct object slot

dispreference at the classes **someone** and **asset**. The argument heads under these classes on the whole occurred because of (i) parser errors, for example *someone*, and (ii) lexical ambiguity, for example *venture*, which occurs under both **activity** and **asset**. The acquired preferences were intuitive, however the precision score was reduced because no preferences were provided in LDOCE for *begin*.

There were also many cases where preferences were attested in the corpus data, but did not match the ones in LDOCE. Some genuine FPs arose because the acquired preferences were faulty. For example, a preference for **body part** at the direct object slot of **devise** arose because of lexical ambiguity. Other FPs arose because of errors in the mapping, rather than any fault in the preference acquisition system. For example, the direct object slot of *drop* included a preference for the WordNet class **piece**. This should be covered by the LDOCE preference **abstract**, but was omitted from the mapping. Other acquired preferences appeared legitimate, but were simply not recorded in LDOCE. For example, LDOCE provides preferences of **abstract** and **human** for the direct object slot of *like*. The acquired preferences however included **artifact**, **substance** and **natural object**, which were not covered by the LDOCE preferences. The fault appeared to be with LDOCE in this case.

WSD, particularly FirstS, concentrated areas of preference, so that preferences were detected more readily. This increased recall but also reduced precision. The reduction in precision is presumably accounted for by the fact that the lexicographers are less likely to identify weaker preferences, and preferences for less strongly selecting verbs. Consequently, it is more likely that there will be a mismatch between LDOCE and the acquired preferences.

4.4.2 CIDE Evaluation

We performed a token-based evaluation using the definition examples in CIDE (Procter, 1995). This experiment was done on a small scale since it required manual sense tagging of the argument heads in the CIDE definitions. A sample of ten verbs were selected from the sample of 30 verbs introduced in chapter 3 on page 65. The evaluation was performed on the TCMs acquired for these ten verbs using data from Lexicon A.

The dictionary, as with most others, provides alongside each entry a list of example uses. From these the head nouns in the subject and direct object slots, and those in the noun phrase of PPs,

Table 4.5: Effect of WSD on proportion of verbs with acquired preferences

WSD	% verbs cut at root	% verbs with prefs in ATCM and none in LDOCE
none	25	4.2
SPass	22	4.6
FirstS	18	4.9
COMB	18	4.9

were extracted manually. Each noun in the specified slot was assigned the WordNet class that best represented the sense of the noun. As much as possible, senses were selected which contained the noun as a direct member, i.e. one of their synonyms. In cases where an appropriate class could not be found from the senses of the noun, a class was selected with a meaning as close as possible to the meaning of the head noun. For the PP slot, clear cut cases of phrasal verbs were ignored since these are handled separately from PPs by the SCF acquisition system.

The acquired preferences were evaluated using the sense tokens, obtained manually from CIDE, for the specified slot. Each token was scored correctly if it fell at or under one of the preferences (again using a threshold of 1 on the association score to establish which classes exhibit a positive preference). A simple system which stated that there was a preference for all the WordNet roots, or all the classes in any cut across WordNet, would have attained 100% accuracy. A baseline was used which considered the proportion of classes on the TCM with preferences above the threshold for each token. This is defined in equation 4.2 where i is a token from the test sample S, and v is the verb specified in the instance i.

$$\frac{\sum_{i \in S} \frac{|\text{classes in TCM}_{v_i} \text{ with score above threshold}|}{|\text{classes in TCM}_{v_i}|}}{|S|} \tag{4.2}$$

The baseline did not explicitly account for the specificity of the cut, although the specificity of the TCM did affect the baseline since a more specific cut typically comprised more classes, with less of these over the threshold, (those that were over the threshold were typically stronger). The baseline was intended to reflect the chance that an item fell under one of the classes on the TCM that expressed a preference, if the classes showing a preference were picked at random. The baseline assumed that the token senses were distributed uniformally under the 11 WordNet roots. The tokens would not have been distributed uniformally, however they would be reasonably spread, since they were taken from the examples for a number of verbs.

Figure 4.4 illustrates the evaluation process for one token using two different ATCMs. The tuple $< believe, direct\ object, robber >$ was observed in the dictionary. This token was manually assigned the class **person**. The token fell under both the ATCMs, with and without FirstS. However, FirstS produced a more specific ATCM. Without FirstS, the baseline ratio for this instance was $\frac{12}{25}$ because there were 12 classes on the TCM above threshold, out of 25. With FirstS, the ratio was $\frac{9}{26}$.

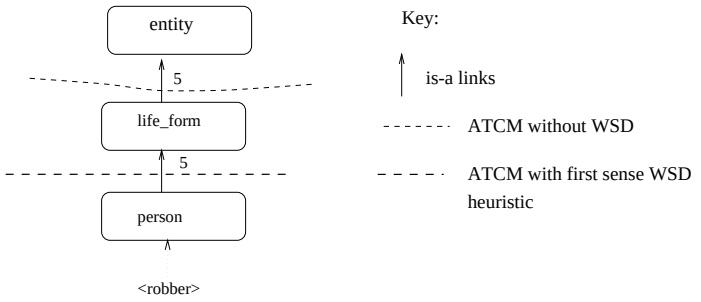

Figure 4.4: *Robber* under ATCMs for *believe* object slot

Table 4.6: CIDE evaluation for ATCMs

slot	WSD	% correct	BL	diff
obj	none	92	58	34
obj	SPass	75	37	38
obj	FirstS	89	48	41
obj	COMB	80	35	45
subj	none	88	67	21
subj	SPass	74	42	32
subj	FirstS	83	47	36
subj	COMB	76	35	41
pp	none	88	65	23
pp	SPass	63	41	22
pp	FirstS	85	62	23
pp	COMB	68	47	21

Table 4.7: CIDE evaluation, object slot

model	%correct	BL	diff
ATCM	92	58	34
LLRTCM	71	25	46
PTCM	34	3.5	30.5
prior	83	27	56

Results are displayed in table 4.6. The % correct was calculated over all items where preferences were acquired, because they occurred in lexicon A with a frequency above the threshold (as before, this was set at 9). The last column shows the difference between the % correct and the baseline (BL). For object and subject slot position this difference increased from (i) not performing any disambiguation, to (ii) the iterative technique alone, to (iii) using the first sense heuristic to (iv) the combination of the first sense heuristic and the iterative approach. The larger % correct score in the case without WSD was due to a larger proportion of classes with a preference above the threshold on the cut. This made it more likely that a test item was subsumed by a class with a positive preference. The larger difference between % correct and the baseline with WSD was because WSD produced more specific cuts. These tended to have a lower baseline ratio since there were less areas of preference over a larger number of classes. The preferences were more discriminatory.

For the PP slot, no improvement in performance was associated with WSD of the input data. Indeed the iterative approach seemed to be slightly detrimental. This was due to the need to specify both the verb and preposition for acquisition and application of these preferences, and consequent shortage of data for training. The sparse data problems were highlighted by the fact that over half of the instances from the dictionary could not be used as the verb and preposition had not been seen with sufficient frequency (greater than a threshold of 9) in lexicon A.

Results for the ATCMs at the object slot are contrasted with the results for the PTCMs and LLRTCMs, all without WSD, in table 4.7. The last column displays the difference between % correct and the baseline (BL). The difference to the baseline highlights the discriminatory power of the preferences. This difference was largest for the LLRTCMs.

4.4.3 Task-Based Evaluation - WSD

The type and token evaluations required the use of a threshold on the acquired preferences. This was because the LDOCE constraints are supplied as restrictions, and are not given a score on a continuous scale. The examples supplied in CIDE were also interpreted on an 'all or nothing' basis. There was no notion of the likelihood of the examples. These evaluations allowed us to compare acquired information with that specified by humans a priori. In addition to these evaluations, we examined how our preferences performed on NLP tasks.

In a task-based exercise, a threshold is not required. However, a threshold can be used to constrain the application of the preferences, only ranking items which have preferences above the threshold. For the WSD evaluation we used a threshold of 1 on the association scores of the ATCM

before applying them.

In this subsection, we describe the results obtained when we applied our TCMs to the task of WSD. We described this application and gave some results in chapter 3. In that chapter, WSD with preferences was used to reduce the noise in the input data to the selectional preference acquisition system. In this chapter, we report the results obtained when we used the WSD task for evaluating the performance of the different model types and various parameter settings of our system. We also compare the results obtained to those reported in the literature for application of acquired preferences to WSD.

The WSD task involved finding the correct sense of each target lemma. The senses were ranked according to the scores of any of their superordinates that featured on the TCMs. This process was illustrated in figure 3.5 on page 64. For a target noun, the sense was selected which had the highest preference score. The preference score for a sense was inherited from its superordinate classes on the TCM.

The WSD evaluation was performed on the SemCor data, since this is freely available and, consequently, has frequently been used by other researchers. Additionally, we used the SENSEVAL test suite for ensuring our preferences were performing reasonably compared to other systems using selectional preferences.

For the SemCor evaluation, we obtained ATCMs for the sample of 30 verbs referred to on page 65 of chapter 3. The ATCMs were obtained by training on the data in lexicon A. The preferences were evaluated with each of the WSD options applied to each of the three target slots. We compared the performance of the ATCMs at the direct object slot without WSD to those produced using the original method of pruning WordNet (Abe & Li, 1996), and to those produced using data specific to a SCF. We also used a smaller portion of the data to evaluate the ATCMs with and without proper noun recognition.

Precision was calculated as the number of instances where the correct answer was included in the classes returned by the system, divided by the number of instances for which the system attempted disambiguation. Recall was calculated as the number of instances where the correct answer was included in the classes returned by the system, divided by the total number of test instances. Systems which do not rely on sense tagged instances for training have typically used the random baseline for comparison (Kilgarriff et al., 1998). This is $\frac{1}{|senses|}$ averaged over all instances in the sample. We devised a higher baseline, the MCBL defined on page 65, to allow for the fact that our system could choose more than one sense tag. For this baseline, each target instance $\frac{1}{|senses|}$ was multiplied by the number of classes left over after disambiguation (thus a baseline for precision only). The MCBL baseline did not indicate the semantic proximity of the senses remaining after disambiguation and was consequently rather high. Comparing our results to this baseline did not allow for cases where our system was settling on related senses.

Table 4.8 displays the results for the object, subject and PP slots using the sample of 30 verbs and training data from lexicon A. All figures given are percentages. The differences between the WSD options were not significant. The main effect of WSD was observed in the previous chapter: it produced more specific cuts and reduced the number of root cuts. For the subject and object slots this did not make a difference, as for these all 30 verbs have cuts below the root. Selectional preferences acquired for the PP slot performed poorly on the task of WSD. This was presumably

Table 4.8: SemCor evaluation

slot	WSD	recall	precision	MCBL	RBL
obj	none	35	48	43	28
obj	SPass	28	43	37	26
obj	FirstS	31	42	42	28
obj	COMB	28	41	39	27
subj	none	35	51	47	27
subj	SPass	31	48	45	26
subj	FirstS	35	52	47	27
subj	COMB	35	53	47	27
pp	none	9	27	40	26
pp	SPass	9	29	39	25
pp	FirstS	15	43	40	26
pp	COMB	13	43	39	24

affected by weak selectional properties of the 30 verbs for the head noun in the NP within the PP. Additionally, the poor performance was exacerbated by sparse data problems. There was considerably less training data available because the selectional preferences were specific to the preposition as well as the slot. In this case, the FirstS option did increase precision, although the difference was not significant.

We used the same data for comparing the ATCMs produced by our system (no pruning) and those produced using the strategy of pruning WordNet devised by Li & Abe (Li & Abe, 1995; Abe & Li, 1996). The comparison is shown in the first two rows of table 4.9, for ATCMs obtained for the object slot without WSD. The pruned version had a higher precision because the items attempted were easier, this was reflected by a higher baseline. The differences between the precision and baselines were not significant. The main difference was in the number of root cuts. Li & Abe style pruning caused more root cuts, because preferences were less readily observed. The SCF row of this table shows the results obtained when the training data was specific to a SCF. For this experiment, we used the frame `[np v np]` and contrasted it with the results obtained using the data at all direct object slots. There was very little difference. The benefits of being specific to the `[np v np]` SCF were lost by the reduction in data available for this SCF.

We also obtained ATCMs for the same verb sample, using data from 1.8 million words of parsed text from the BNC. This was a portion of the same data used to construct lexicon A. Two lexicons were built, one using GATE (Cunningham et al., 1995) named entity recognition (lexicon B) and one without proper noun recognition, but which used pattern matching to detect dates and monetary amounts (lexicon C). The ATCMs for the target set of 30 words were contrasted on the SemCor WSD task. The results are displayed in the lower part of table 4.9. Proper noun recognition increased the coverage of the data by reducing the number of verbs with root cuts. Precision and the precision baseline were consequently lower since the preferences were more discriminatory and so returned fewer sense labels. Although proper noun recognition did increase

Table 4.9: SemCor evaluation - ATCMs direct object slot

	recall	precision	MCBL	RBL	% verbs cut at the root
No pruning	35	48	43	28	0
L&A pruning	34	52	51	28	20
SCF specific	36	48	44	28	0
	Training on 1.8 M words				
Ignore proper nouns	29	55	45	29	43
PN recognition	31	48	42	28	25

Table 4.10: SemCor evaluation - sample of 395 verbs

model type	WSD	recall	precision	MCBL	RBL
ATCM	none	32	48	44	26
ATCM	SPass	34	48	44	26
ATCM	FirstS	33	46	43	26
ATCM	COMB	33	47	43	26
LLRTCM	none	35	49	42	26
PTCM	none	35	48	46	26

coverage and provide more discriminatory preferences, we did not carry it forward for further experiments because of the lengthy computation required for large quantities of data.

We also obtained results for TCMs from 395 verbs. taken from a random sample of 500 sentences. [3] The 395 verbs all occurred in lexicon A with frames involving a direct object slot. Performance on the WSD task was compared for the three model types, ATCM, LLRTCM and PTCM, acquired for the direct object slot without WSD on the input data. The TCMs all obtained precisions in excess of MCBL, the largest difference was for the LLRTCM model. WSD of the input data did not improve performance of the preferences on the WSD task.

These results are hard to compare directly with other researchers evaluating on the SemCor data because of experimental differences. Resnik (1997) obtained an average of 44.3% accuracy for object and 40.8% for subject slot using preferences acquired for 100 of the most strongly selecting verbs. Resnik did not allow the assignment of multiple tags, but selected at random from the senses returned by the system when more than one tag was suggested by the preferences. Abney & Light (1999) obtained scores using the same training and test data as Resnik. Abney & Light also selected randomly between multiple senses returned by the system. They obtained an accuracy for the direct object relationship of 42%. Accuracy increased to 54% when they used preferences obtained from training on the full set of the BNC. We disambiguated the direct objects which occurred with 395 randomly selected verbs, rather than verbs with strong preferences. Verbs selected at random were expected to perform worse than verbs chosen because of their se-

[3] This is the same set of verbs referred to on page 36 in chapter 2.

lectional properties. The test data was therefore more difficult and performance was anticipated to be adversely affected by this. However, this was not the only difference between our experiment and those of Resnik and Abney & Light which was predicted to affect performance. We permitted multiple assignment of sense tags and so in this respect our task was easier. The MCBL to some extent took the possibility of multiple sense selections into account, although this was rather a stringent baseline. It did not allow for the fact that the TCMs made multiple sense selections only where these occurred in the same vicinity of WordNet.

SENSEVAL permitted multiple sense selection. This was done by applying a probability distribution over the senses returned for each test item. If no probability distribution was supplied then a uniform one was used over the sense tags returned. Our system performed comparably with the other system (OTTAWA) which used only automatically acquired preferences for disambiguation (Kilgarriff & Rosenzweig, 2000). Performance of our system on the coarse grained noun task was 69% precision with 20% recall. These results are comparable to those of the OTTAWA system which obtained 71% precision and 8% recall (Kilgarriff et al., 1998). The random baseline, with multi-word identification performed at 58% precision and 58% recall on this task.

4.4.4 Task Based Evaluation - Pseudo Disambiguation

For this task, <verb1 noun verb2> tuples were created where the pair verb1:noun had occurred in the test data, but verb2:noun had not. The TCMs had to identify which verb was more likely to occur with the noun.

We evaluated our TCMs on this task since it bears some relation to a structural disambiguation task. The preference models were used to find the most likely combination between two pairs of words. In the pseudo-disambiguation task, word pairs are artificially created and the system is expected to prefer the genuine pairs, for a given slot, to the artificially created ones. In structural disambiguation proper, the decision is made between argument heads at different slots, using preferences specific to the attachment site, for example of the verb and NP for PP resolution. Those that have evaluated acquired selectional preferences on a structural disambiguation task have typically used supervised training data (Abe & Li, 1996; Resnik, 1993a). A structural disambiguation exercise using unsupervised training data was possible, but direct comparison with a supervised approach would not have been appropriate. The advantage of doing the pseudo-disambiguation task was that supervised training data was not required. For structural disambiguation, the output from the shallow parser could be used for training. However, there were likely to be errors on PP attachment which would make comparison difficult with systems using semi-automatically parsed text, like that of the Penn Treebank. The task has also been referred to as pseudo-WSD (Dagan, Marcus, & Markovitch, 1993; Lee, 1997) since the two verbs can be thought of as two senses of a pseudo-word formed by combining the two verb forms.

We used the data in lexicon A as training data, and data from a further portion of 8 million words of parsed text from the BNC as test data. Tuples were selected such that the combination of the noun with either verb1 or verb2 did not occur in the training data. However, the noun in isolation had occurred in the training data. Verb1 was required to have occurred in the test corpus with the noun. Verb2 meanwhile was selected at random according to its frequency distribution in the test data. The system decided which verb was more likely to have been seen with the noun in

Table 4.11: Pseudo-disambiguation evaluation

Model	WSD	Precision	Recall
ATCM	none	58	55
ATCM	SPass	55	51
ATCM	FirstS	59	55
ATCM	COMB	59	54
PTCM	none	58	58
PTCM	SPass	55	53
PTCM	FirstS	56	56
PTCM	COMB	57	55
LLRTCM	none	58	58
LLRTCM	SPass	59	59
LLRTCM	FirstS	59	59
LLRTCM	COMB	59	59

the test data on the basis of the preference scores. Precision and recall were calculated from these decisions. Precision was calculated as the number of times where the system decided correctly, divided by the number of attempts at a decision. Recall was calculated as the number of times where the system decided correctly, divided by the total number of instances. Precision and recall were distinct in this experiment since our system did not make a decision in cases where an item was not covered by the TCM, or where the score was the same for both verbs. The lower bound for this experiment was 50%, for both precision and recall. This would be expected if the system made a random decision. The upper bound was below 100%, but has not been determined. A recall of 100% cannot be expected because the false pairs are generated artificially. It was quite possible that, on occasions, they were in fact more plausible than the pair actually attested in the corpus. To determine an upper bound it would have been necessary for lexicographers to decide which pair was more likely, without recourse to the correct answer. A measure of inter-lexicographer agreement would also have been required. An upper bound was not determined because of the substantial human effort required, we simply noted that the system was expected to perform below the 100% level.

For the experiment, we used the selectional preferences obtained for verbs found in the sample of 500 hand-parsed sentences with sufficient direct objects (again, the threshold was set at 10 or more) for preference acquisition. We randomly produced tuples according to the procedure outlined above. Precision and recall figures for the ATCMs, PTCMs and LLRTCMs are provided in table 4.11. The model types are displayed in this table with the different WSD options. All experiments were conducted using preferences and data from the direct object slot. The scores for the ATCMs and LLRTCMs improved slightly with the first sense heuristic, whilst the PTCMs were slightly worse off with WSD. The WSD options did not give significant differences in terms of precision and recall.

The differences in performance of the different model types and WSD options were not signif-

icant. There was a large disparity between our results on this task and those obtained by Rooth et al. (1999) and Pereira et al. (1993). They obtained results with near 80% accuracy (their system was allowed to decide in all cases). These were far superior to ours. It is likely that some of the difference was due to the difference in the size of the training portion. We used the data in lexicon A, which is built from 10.8 million words of the BNC, whereas Rooth et al. used the entire BNC (90 million words) for training. Pereira et al. used a 44 million word corpus for training. Abney & Light (1999) demonstrate a significant improvement on the WSD test when increasing the size of the training corpus. (Rooth et al., 1999) also restricted the verb and noun lemmas to be ones which occurred between 30 and 3000 times in a specific relationship in the training corpus. Pereira et al. used verbs with a frequency between 500 and 5000. We placed no frequency threshold on the lemmas, but the verb1:noun pairs and verb2 were selected according to their frequency distribution.

The substantial difference in performance of our system to the ones using automatically produced classifications may not be solely attributed to differences in the training data. Li & Abe (1996) demonstrated that automatically clustered models were more accurate at the task of structural disambiguation than models obtained as tree cuts in WordNet. WordNet, meanwhile, permitted better coverage of the data.

4.5 Conclusions

In this chapter, we evaluated our ATCM, PTCM and LLRTCMs, with the various WSD options. We used both type and token-based evaluation methods, and also performed two task-based evaluations, used by others in the literature.

The acquired preferences covered a considerable portion of the LDOCE restrictions, whilst also including preferences which were not recorded by the LDOCE lexicographers but were quite intuitive and plausible. A significant source of error for this type-based evaluation arose from the difficulty in mapping between LDOCE and WordNet.

Manually tagged examples provided in the verbal entries of CIDE were used for the token-based evaluation. Since the tokens were not collected from corpus data it was again not expected that all examples would be covered. This evaluation favoured the more discriminatory models, those being the LLRTCMs and those incorporating WSD, particularly FirstS. These models provided a relatively large difference between the CIDE examples that they covered and the random baseline.

The type and token-based evaluation necessitated the use of a threshold to determine coverage of (i) the LDOCE restrictions and (ii) the CIDE examples. Our TCMs expressed preference on a continuum, but this was lost in these evaluations. The choice of threshold affected performance for the LDOCE evaluation, with a higher threshold increasing precision whilst reducing recall. The task-based evaluations did not require a threshold.

Our preference models performed similarly on the WSD task to those of others. Performance on the pseudo-disambiguation evaluation, however, was considerably lower than that achieved by some of the distributionally based classifications reported in the literature. At least some of the discrepancy was probably due to differences in the training and test data. Interestingly, there are no other researchers who have used this evaluation for preferences acquired automatically using a manmade taxonomy, such as WordNet. Li & Abe (Li & Abe, 1996) indicated that higher levels

of precision can be obtained using automatically constructed models for structural disambiguation. Presumably automatically constructed taxonomies fit naturally occurring corpus data more accurately. Comparisons of models using automatically constructed semantic classes with models using manmade resources on the same test suite are needed.

The different model types produced slightly different results from each other, although on the whole these differences were not significant. The ATCM and PTCM models performed similarly on the LDOCE and CIDE evaluations. The LLRTCMs showed the largest difference to the baselines on both the CIDE and the WSD evaluations. The WSD options did produce slightly better results in some experiments. For example FirstS significantly increased precision in the face of sparse data at the PP slot. However, the results did not show consistent improvement with WSD of the input data. The main affect of WSD was to increase the coverage by increasing the specificity of the TCMs, thereby reducing the number of root cuts.

Using proper nouns provided more discriminatory preferences and increased coverage. But, given the need to process large quantities of data, we did not pursue this for diathesis alternation detection. Obtaining the data specific to the SCF did not significantly improve or degrade performance. The reduction in noise was accompanied by a reduction in the training data. For diathesis alternation detection we required preferences specific to SCFs. At least the reduction in training data was compensated for by the reduction in noise.

Chapter 5

Identifying Diathesis Alternations

5.1 Introduction

This chapter concerns the automatic identification of diathesis alternations. Diathesis alternations are different ways in which the arguments of a verb are expressed syntactically. They are sometimes accompanied by slight changes in the meaning of the verb. An example of the causative-inchoative alternation is given by the sentences in example 13 below. In this alternation, the object of the transitive SCF can also appear as the subject of the intransitive SCF.

(13) a. The boy broke the window.

b. The window broke.

We are specifically concerned with alternations involving NP and PP constituents, since our selectional preferences can only be applied to these slots. Alternations involving NPs and PPs can be broadly divided into three categories according to their syntactic behaviour. Firstly, those in which arguments are optional, and are omitted in one realization, for example the unspecified object alternation shown in (14) below. Secondly, those in which particular argument types occur in different slots with different grammatical roles in the alternate frames, an example of this is the dative alternation shown in 15. In (15a). the argument acting as the 'recipient' (*the dog*) occurs as the direct object in the double object construction. Whereas in (15b) the recipient appears in the prepositional phrase. In this example, the argument acting as the 'theme' (*a bone*) occurs as the indirect object (second object) of (a) but the direct object (first object) of (b). We refer to these as 'role switching' alternations (RSAs). And thirdly, those involving both omitted arguments and role switches, for example the causative-inchoative exemplified above in example 13. In this thesis, we are particularly concerned with RSAs (with or without omitted arguments). In our experiments, we have used the syntactic and semantic evidence gleaned automatically from corpus data. We looked for cases where semantically similar argument heads appeared in different slots in the alternating syntactic realizations.

(14) a. The boy ate the popcorn.

b. The boy ate.

(15) a. She gave the dog a bone.

b. She gave a bone to the dog.

The next section (5.2) briefly introduces a little of the background literature on alternations and some issues relating to their computational treatment. Section 5.3 motivates the need for automatic identification of verbal participation. We emphasize why alternations are interesting from a theoretical perspective and what practical uses can be made of them. In section 5.4, we look at related work on the acquisition of alternations and point out the difficulties involved. We outline the approaches we took for automatic identification of alternations in section 5.5. Our methods combine automatically produced SCFs and preferences. Levin (1993) provides a comprehensive list of alternations, alongside a manually produced classification of verbs according to their participation in these alternations. We define the scope of our experiments in terms of this classification in section 5.6. We then look at automatically acquired SCF data in section 5.7 to show the alternations which we were able to examine because we had sufficient data. Section 5.8 contains our experimental results using our automatic methods and 5.9 contains a summary and our conclusions.

5.2 Some Background on Diathesis Alternations

Alternations relate to both the syntax and semantics of natural language. Transformational grammarians have studied alternations from a syntactic perspective. They have investigated the ways in which the same underlying thematic role is expressed in surface position (Radford, 1989; Fillmore, 1970). Radford took the view that the same set of selectional restrictions should apply in all surface positions of the same underlying thematic role. There are a number of problems with this. It is not easy to manually identify selectional restrictions in terms of abstract classes, without explicitly listing the nouns themselves (Fillmore, 1970). Also there are many cases where participation depends not only on the verb, but on the argument too. Verbs which alternate do not do so for all argument heads (Montemagni, Pirrelli, & Ruimy, 1995). In example (16) below, the transitive variant of the causative-inchoative alternation for *ring* is unlikely to occur in a corpus of English with the argument *alarm clock* in the object slot. Meanwhile, the intransitive variant is quite plausible.

(16) a. *The boy rang the alarm clock.

b. The alarm clock rang.

Thus, a particular combination of argument and predicate may preclude participation. Furthermore, some predicates will not participate in an alternation at all, even though they occur with one of the required syntactic forms. Thus, alternations are not fully productive. They were described as *semi*-productive by Briscoe & Copestake (1996) who argued that it is not feasible to enumerate the conditions under which alternations apply when handling them in a computational lexicon. Sanfilippo (1994) used MRDs for determining participation for specific lexical entries. However, MRDs are manmade resources, which are open to human error and are not tailored for any particular corpus. MRDs also lack the frequency information that is available when acquiring

lexical information from corpora. Certainly, corpus evidence plays an important role in the study of productivity issues and the interaction of argument and verb on verbal participation.

Alternations have often been described with lexical rules in lexicalist grammar formalisms, such as Head-Driven Phrase Structure Grammar (HPSG) (Pollard & Sag, 1987, 1994). Lexical rules have an input (base form) and output (derived form). Briscoe & Copestake (1996) looked at semi-productive lexical rules generally, including sense extensions such as **vehicle-name** → **verb-of-motion** as well as verb alternations. They advocated the use of probabilities to help control the application of the lexical rules. Probabilities were used for productivity estimates to allow for the semi-productive nature of lexical rules. The probabilities were recommended to prevent lexical rules being applied in all cases having the base form, and to predict the likelihood of a previously unseen case. The motivation behind using probabilities was in line with the Gricean Maxim of Manner: there is an implicit understanding between speakers not to use rare or abnormal forms. Users of a language will more readily produce and expect more commonly used forms. The paper illustrated how probabilities for constraining the application of lexical rules might be estimated. This was done using attested evidence for the productivity of a lexical rule to help provide a probability estimate for unseen forms. Crucially the method relied on observing some occurrences of the lexical rule and estimating the probabilities from these observations.

The productivity of a lexical rule $A \rightarrow B$ was calculated by Briscoe and Copestake as :

$$\text{Prod}(rule) = \frac{\text{num}(B)}{\text{num}(A)} \tag{5.1}$$

This was a type ratio, for lexical entries, where num(A) was the number of attested lexical entries which matched the input, or base form, of the lexical rule and num(B) was the number of entries observed to match the output, or derived form. Montemagni & Pirrelli (1995) showed that, for the causative-inchoative alternation at least, there does not seem to be a unique direction between base and derived forms for all verbs. In this thesis, we remain neutral to the issue of directionality of the alternations and do not rely on the notion of a base entry.

There is a widely held view that the meaning of a verb and its participation in alternations are connected (Pinker, 1989; Jackendoff, 1990; Levin, 1993; Levin & Rappaport Hovav, 1996). However, the exact aspects of meaning which give rise to the syntactic behaviour are subtle and often rather elusive. Pinker (1989) and Jackendoff (1990) looked at alternation phenomena from a semantic perspective. They were interested in how the syntactic realisations of a predicate are determined by its meanings. Their approaches hinged on linking rules which mapped semantics to syntactic realizations.

Jackendoff did this using what he termed 'lexical conceptual structures' (LCSs) as the basis for defining semantic behaviour. In this framework, basic semantic categories such as **Thing**, **Event** and **State** were combined using formation rules. Thematic roles were an important part of this machinery. They tied the argument positions in a LCS to the NPs in the syntactic expression. A mapping was established between LCSs and the syntactic argument structures. Verbs with similar meanings could occur in the same sorts of LCSs. These LCSs were in turn realised as syntactic structures characteristic of the verbs. The framework included notation to allow alternating forms to be related at the LCS level.

Pinker looked at argument structure from a learning perspective. He looked at the way children learn generalisations, construed as lexical rules, so as to produce new argument structures in lan-

guage. The argument structures of verbs were said to be projected from the underlying semantic structure of the verb via linking rules. Pinker stipulated that the lexical rules act on the semantic structures, producing a change in meaning. The change in semantic structure gives rise to a corresponding change in argument structure. Semi-productivity occurs because, for some verbs, the semantic change does not tally well with the original semantic structure of the verb, and therefore the lexical rule cannot be applied. He proposed a common semantic component ('a thematic core') which is responsible for a group of verbs behaving in the same way.

The link between the semantics and syntax of alternations is of great interest. If some semantic components can be identified with syntactic behaviour then this opens up the possibility of learning something about the semantics of an unknown word by observing its syntactic behaviour. Conversely, if one knows enough about the semantics of a verb then one can generate new forms, regardless of whether they have been seen before or not. However, identifying these semantic components is far from straightforward (Levin & Rappaport Hovav, 1996).

Levin (1993) consolidated diathesis alternation research by designing a framework which encompasses both the syntactic behaviour and the underlying semantics. She has produced a list of alternations involving NP and PP constituents and has manually classified over 3000 verbs according to their participation in these alternations. Alternations involving other complement types, such as sentential complements, are not included. The verb class taxonomy features 191 verb classes and provides the key semantic and syntactic characteristics of each class. The verb classes show considerable semantic cohesion, providing evidence for the link between syntactic behaviour and meaning. Levin's classification is extensive enough for practical use and is now used by a wide number of NLP researchers (Dorr & Jones, 1996; Dang, Kipper, Palmer, & Rosensweig, 1998; McCarthy & Korhonen, 1998; Stevenson & Merlo, 1999; McCarthy, 2000).

Until recently, the NLP research concerned with alternations has concentrated on issues of representation of lexical rules (Sanfilippo, 1996, 1994; Briscoe & Copestake, 1999; Bredenkamp, Markantonatou, & Sadler, 1996), productivity issues (Pirrelli, Ruimy, & Montemagni, 1994; Montemagni et al., 1995; Montemagni & Pirrelli, 1995; Briscoe & Copestake, 1996) and cross-linguistic studies (Pirrelli et al., 1994; Nicholls, 1994, 1995). This thesis is concerned with the automatic identification of RSAs. In the next section, we consider why such information is useful for NLP. Before moving on, we note that theoretical linguistic research on alternations could also benefit from the use of tools capable of automatically suggesting new participants from corpora. Identifying new participants might help narrow the search for the components of meaning that drive the syntactic behaviour, and for the necessary and sufficient conditions for the alternating forms.

5.3 Motivation

In this section, we describe some of the ways in which alternations have been used for NLP purposes. These demonstrate the need for information concerning alternations to be stored in a computational lexicon.

Diathesis alternations have been suggested as a basis for improving lexical acquisition (Ribas, 1995a; Korhonen, 1997; Briscoe & Carroll, 1997). Korhonen (1997) proposed using diathesis alternations in SCF acquisition. The goal of her work was to improve the statistical filter of the

SCF acquisition system of Briscoe & Carroll (1997). This is the same SCF acquisition system that we have used in this thesis. The statistical filter was intended to decide whether a SCF observed for a particular verb was genuine or not. This was necessary because a SCF may be detected with a particular verb because of noise, for example arising from parser errors. To see how diathesis alternations might improve the SCF acquisition process we need to delve into the details of the statistical filter.

In the original system (Briscoe & Carroll, 1997), the filter works using hypothesis testing on binomial frequency data. This is based on the binomial filter devised by Brent (1993) for SCF acquisition. The observations of verbs occurring with SCF classes in the corpus are construed as a binomial frequency distributions. Binomial distributions are usually exemplified using a number of coin flips. The outcome of each flip is one of two alternatives (heads or tails). The number of flips (n) is fixed. The probability of a particular outcome with probability p occurring m times out of n trials is given by the binomial distribution in equation 5.2:

$$P(m,n,p) = \frac{n!}{m!(n-m)!} p^m (1-p)^{n-m} \tag{5.2}$$

The probability of the event happening m or more times is:

$$P(m+,n,p) = \sum_{i=m}^{n} P(i,n,p) \tag{5.3}$$

In the SCF acquisition scenario, the trials are the occurrences of a verb (v) in the corpus used for acquisition. The outcome of the trial is either an occurrence of v with a particular SCF (i), or an occurrence with some other SCF. The probability $p(v\neg i)$ is the probability that v does not legitimately occur with i, yet is observed with i in the corpus. $P(m,n,p(v\neg i))$ is the probability that SCF i was seen m times with v in a corpus having n occurrences of v, when this was due to noise. This can be used in hypothesis testing to see if the occurrence of i with v has occurred more than would be expected by chance. A threshold is set on the value $P(m+,n,p(v\neg i))$, usually less than or equal to 0.05, to give a confidence level that sufficient occurrences of i have been observed with v for this to be a genuine SCF for v. The null hypothesis is that v does not legitimately occur with i. The alternative hypothesis is that v does occur with frame i. The null hypothesis is rejected if the number of occurrences of i with v exceed the threshold determined by $P(m+,n,p(v\neg i))$.

An estimate is required for $p(v\neg i)$. Briscoe & Carroll estimate $p(v\neg i)$ using information from the ANLTdictionary and from the Susanne corpus (Sampson, 1995). The calculation is shown in equation 5.4. The estimate is dependent on SCF, but is irrespective of the verb.

$$p(v\neg i) = (1 - \frac{|\text{ANLT } verbs \text{ in class } i|}{|\text{ANLT } verbs|}) \frac{|patterns \text{ for } i|}{|patterns|} \tag{5.4}$$

The first term estimates the probability of a verb not occurring with SCF i. The second term estimates the probability of the syntactic pattern associated with SCF i occurring.

Korhonen (1997) used diathesis alternations to improve the filtering process. The basic idea was to make use of correlations between SCFs because of diathesis alternations. When acquiring

SCFs automatically, one does not know before hand the correlations for a specific verb. However, one might be prepared to use a guess at the correlations from all verbs taken collectively. Korhonen did this using the SCF entries in the ANLT dictionary. Nine alternations from the linguistic literature were used, and also other rules were automatically identified by finding correlations between the SCFs in ANLT. The rules were directional rules of the form:

$$\text{SCF class A} \rightarrow \text{SCF class B}$$

One such rule was for the intransitive class (22), with the subcategorization frame [`np v`], alternating with the transitive class (24), which has the subcategorization frame [`np v np`].

$$\text{class 22} \rightarrow \text{class 24}$$

Probabilities were estimated for each alternation rule using the number of verb types in both A and B divided by the number in A. These probabilities were used to alter the value of the threshold for SCF A which was used for hypothesis testing. If a SCF did not occur with a verb more than this threshold, then the occurrence of the SCF with this verb was assumed to have arisen because of noise. The alternation probability was used to lower the threshold for the SCF on the LHS in cases where both SCFs were evident. This was done using the formula in equation 5.5. If the SCF on the left hand side of the alternation rule occurred with a verb, but the SCF on the right hand side did not, then equation 5.6 was used to increase the threshold for the SCF on the LHS.

$$\text{New Threshold} = P(m+, n, p(v\neg i)) \times (1 - \mathit{AlternationProbab}) \tag{5.5}$$

$$\text{New Threshold} = P(m+, n, p(v\neg i)) \times (1 + \mathit{AlternationProbab}) \tag{5.6}$$

There were a number of parameters in Korhonen's experiments, but overall the the alternation probabilities were shown to improve the performance of the system. In subsequent work (1998), Korhonen used alternations, again taken using the SCF entries in ANLT, to directly alter the probability of a SCF given a verb. Identifying alternations from a corpus, rather than from the SCFs entries in ANLT, might improve SCF acquisition further. Verb specific evidence would be useful in determining the entries for an individual verb. Evidence collected over the entire set of verbs can be used, alongside attested evidence, when verb specific information is not available.

Alternations can also be used for the lexical acquisition of selectional preferences (Ribas, 1995a). They permit us to relate alternative forms together when collecting the argument heads occupying a particular relationship with the predicate. Ribas looked at the passive alternation for *present*. The passive alternation was selected since this is easily detected, and *present* was selected since it occurs with similar relative frequencies for the two alternating SCFs. This alternation is shown in example 17.

(17) a. The doctor presented the issues.

b. The proposal was presented by the director.

Ribas obtained selectional preferences from the data in the WSJ for the subject and object slots of *present*. He did so in three different experiments. In the first, the passive alternation was ignored. Thus, using our example, *doctor* and *proposal* would have both contributed to the selectional preferences of the *subject* slot. In the second experiment, the passive alternation was detected and argument heads from different semantic roles were separated. Thus only *doctor* from our example would have been used for acquiring subject slot preferences. In the third experiment, the passive alternation was detected and applied. Argument heads having the same functional relationship (semantic role) with the verb were combined. Thus, *doctor* and *director* would have been combined. The preferences acquired from the third experiment increased precision and recall on a WSD task, compared to the first experiment. Simple detection of the passive (experiment 2) reduced precision and recall. This was explained by the reduction in the volume of data available.

Another use of alternations in SCF acquisition is for the recovery of full predicate argument structure (Briscoe & Carroll, 1997). Boguraev & Briscoe (1987) pointed out that participation in alternations can help in determining control options for predicates. They can help classify a verb as equi or raising. For example, a raising verb such as *believe* is a two place predicate. Meanwhile, *persuade* is an equi verb (a three place predicate). They share the SCF associated with example (18a) below. *Believe* also takes (b) whilst *persuade* takes (c). Thus the control classification of (a) is determined by the occurrence of the alternative form.

(18) a. I believed/persuaded Bill to be a good man. [`np v np vp_infinitive`]

b. I believed Bill was a good man. [`np v s_comp`]

c. I persuaded Bill John was a good man. [`np v np s_comp`]

Alternations involving verbal complements are not tackled in this thesis. However, given that there are correlations between alternations (Levin, 1993), techniques that identify alternations involving NPs and PPs might help in establishing participation in other types.

Alternations provide a means of adding generalisations to the lexicon. This is preferable to enumerating cases individually for two reasons. Firstly, it produces a compact lexicon in which connections between alternating forms of the same verb are transparent. Secondly, generalisations over groups of verbs can be readily be observed. Diathesis alternations are typically implemented in the form of lexical rules, as in lexicalist grammar formalisms such as HPSG (Pollard & Sag, 1987, 1994). What is required is a way of determining which verbs participate in which alternations.

The diathesis information required for a verbal entry will vary, depending on the task. At a minimum it should include a reference to the alternations which the verb participates in. Frequency information for the alternating variants might also be useful. This could be used for estimating productivity, in a scheme such as that described by Briscoe & Copestake (1996). Their estimation of productivity, given in equation 5.1 on page 112, rests on the stipulation of the base and derived form for an alternation. In this thesis we do not take up the issue of directionality. Instead we observe data for both alternating variants. We could use our system for determining participation within Briscoe & Copestake's scheme. To do this we might define productivity as the number of verbs where participation is identified, divided by the number of verbs which have at least one of the alternating forms. There are other ways suggested for calculating productivity. Lapata (1999)

estimated productivity using the verb classes listed by Levin (1993). She used the ratio of verbs from a particular Levin class found to alternate to the number of verbs in the class regardless of participation.

Levin's research using diathesis alternations has demonstrated the appeal of alternations as a means for classifying verbs. This large scale classification was not intended to be exhaustive. Extending it with new participants would help others using the classification for their own research. Additionally, the classification lends itself to predicting unseen syntactic behaviour where a new item can be unambiguously classified from seen behaviour. Parallels drawn between the classification and WordNet (Dang et al., 1998) might help in the semantic classification of a new predicate. The appropriate WordNet class of a verb might be determined by first using syntactic evidence to classify it within Levin's taxonomy. Links between this taxonomy and WordNet might then be used to classify the verb in WordNet. Continuing the classification manually would require substantial human effort. Automatic identification of alternations would be a useful tool for automatic, or even semi-automatic, classification.

Different diathesis alternations give different emphasis and nuances of meaning to the same basic content. These subtle changes of meaning are particularly important in natural language generation, according to Stede (1998). Stede considered the changes to the aspectual category (or Actionsart) of a verb using the following four examples: [1]

(19) a. The engine *drained*.

b. He *drained* the engine.

c. The oil *drained* from the engine.

d. He *drained* the oil from the engine.

In his scheme, alternations which affected the denotation (truth conditions) of a verb were termed extensions and treated as directional. He determined a base form of each alternation and application of an alternation added meaning to the base form. In his example 19, (c) was the base meaning denoting the activity, (a) was the resultative reading, which indicated that the engine ends up empty, (d) was introduced by the causative extension and (b) was produced by a combination of the resultative and then causative extensions. These subtle changes in meaning can be exploited in generation provided that the link between the rules and meaning is specified and that the possible alternations for each verb are provided. It is the latter prerequisite that our work addresses.

Thus there are theoretical and practical uses of alternations in linguistics and NLP. Incorporating this sort of information into lexicons has been the subject of previous research (Briscoe & Copestake, 1999; Bredenkamp et al., 1996; Sanfilippo, 1994). Kohl, Jones, Berwick, & Nomura (1998) described how WordNet has been manually supplemented with selectional restrictions and Levin classes. Automatic identification of participants has however only recently received attention. This is the subject of our next section.

[1] These are abbreviated here to remove superfluous text which cannot be contrasted.

5.4 Related Work

In this section we describe methods aimed at classifying verbs according to their participation in alternations. First of all we describe two approaches (Dang et al., 1998; Dorr & Jones, 1996) which extend Levin's classification. These relate indirectly to identification of the verbs involved in particular alternations. We then outline four approaches (Resnik, 1993a; Schulte im Walde, 1998; Stevenson & Merlo, 1999; Lapata, 1999) which, like our approach, classify verbs with reference to corpus evidence.

In Levin's classification a verb can belong to more than one class. This stands to reason as the verbs often have more than one sense. Levin does not explicitly label the sense of a verb when it is listed as a class member with a predefined sense tag, but she lists the verbs with other participants, and the class is typically named using a prototypical member. For example, *strike* is a member of two of Levin's verb classes. It is a member of the **hit verbs** class, along with other members such as *bang*, *beat* and of course *hit*. It is also a member of the **amuse verbs** class, along with other members such as *amuse*, *entertain* and *provoke*. These two classes clearly signify two distinct senses of *strike* illustrated by example 20, where (a) is an instance of the **hit** sense and (b) is an instance of the **amuse** sense.

(20) a. The man struck him with an iron bar.

b. She was struck with the idea.

Dang et al. (1998) modified Levin's classification by adding new classes to make the entire set of classes mutually exclusive i.e. so that there was no overlap between classes. The new classes were termed 'intersective classes' and contained verbs which belong to more than one of Levin's classes. The intersective classes were produced by a procedure that found verbs which appeared in the same set of classes in Levin's taxonomy. Some of these sets of classes became intersective classes. An intersective class was formed where the set of classes contained at least three verbs which occurred in each of the classes in the set. As an additional constraint, the intersective class was only retained provided that it was not subsumed by another intersective class which covered a wider set of classes from the original scheme. Verbs in Levin's index were then reclassified into these intersective classes, provided that they occurred in each of the classes that made up the intersection. On doing this the verb was removed from the original classes that comprised this intersective class. The reclassification had a finer granularity than Levin's scheme and the syntactic behaviour of class members was less diverse. Furthermore, the semantic component of the intersective classes was clearer than in Levin's classification.

Dang et al. demonstrated links between the new intersective classes and the classes in WordNet. Some intersective classes displayed the same sort of distinctions that WordNet subclasses (hyponyms of a common parent) do. For example, two subclasses of *cut*, **separating into bits** and **incision without separation**, were consistent with membership versus non-membership of an intersective class containing the verb *split*. The introduction of WordNet emphasized the semantic relationships between verbs in Levin's scheme. Other intersective classes demonstrated regular sense extensions that are not recorded in WordNet. The intersective classes also allowed Dang et al. to identify cross-linguistic generalisations (English-Portuguese) where translations of verb forms participate in the same alternations.

Dorr & Jones (1996) also highlighted the multiple membership of verb forms in Levin classes. They too wished to show that the relationship between semantics and syntactic behaviour is clearer and more coherent if this is resolved. Instead of increasing the granularity of classes, they increased the granularity of the items being classified, They quantified syntactic behaviour shared by class members and showed that if verb senses were used, rather than verb forms, the shared syntactic behaviour drastically increased.

Dorr & Jones defined the senses for a verb by the Levin classes that the verb belonged to. Thus, each verb sense was defined as a verb form and Levin class combination. For example, the verb *cut* belonged to seven classes and was therefore defined with seven senses. One of these classes was the **split** class, which gave the sense: '*cut*:**split**'.

Dorr & Jones characterised syntactic behaviour using parses of the positive and negative example sentences provided by Levin for each class. These syntactic patterns comprised the major categories within the sentence, including any prepositions. For example, *Tony broke the vase to pieces* provided the pattern [`np v np pp(to)`]. Table 5.1 provides some further examples for three of the classes containing *cut*. The first column provides the class name and the second column lists a couple of members. The third column displays a few of Levin's example sentences and the fourth gives the corresponding syntactic pattern for the example sentence in the third column.

Dorr & Jones conducted two experiments on the shared syntactic behaviour of class members. One was referred to as 'verb-based' and used verb forms and the second was referred to as 'class-based' and used verb senses. For the verb-based experiment, syntactic patterns were collected for each verb from all the classes that the verb belonged to. So for example, all the patterns in the fourth column of table 5.1 would have been included for the verb *cut*. [2] For the class-based experiment, only the patterns from the specified Levin class were collected for a particular verb sense. Thus, only the patterns belonging to the final row [3] would have been collected for the *cut***build** sense.

For the verb-based experiment, any set of syntactic patterns shared by one or more verb form was termed a 'syntactic signature'. The verbs which matched each syntactic signature were placed together in a syntactic grouping. The overlap between the members of this syntactic grouping and the members of each Levin class (a semantic grouping) was calculated. The overlap was the number of overlapping items divided by the average number of items in the Levin class and the syntactic grouping. For the class-based experiment, the same procedure was followed except that the syntactic signatures were obtained using the syntactic patterns collected for verb *senses*, rather than the verb forms. For the verb-based experiment only 6.3% of the 191 Levin classes had complete overlap with the syntactic groupings. In contrast, for the class-based experiment 97.9% of the Levin classes overlapped with the syntactic signatures for the verb senses. These results demonstrated that the syntactic behaviour of the members of Levin's classes is highly correlated with their membership, provided that multiple membership is handled by treating occurrences of the same verb form in several classes separately.

Dorr & Jones went on to propose a method to classify new verbs within Levin's classification. The system used both WordNet and LDOCE. A set of LDOCE grammar codes was specified for

[2] Further examples given by Levin for the three classes in figure 5.1 and the four other classes containing *cut* would have also been included. These are not listed in this table because of the lack of space.

[3] Again, these are not all listed in this table.

Levin Class	some verb members	some of Levin's examples	syntactic pattern
cut	*cut*, *snip*	*Carol cut the bread*	`[np v np]`
		Carol cut at the bread	`[np v pp(at)]`
		**The bread cut*	`*[np v]`
split	*break*, *cut*	*I broke the twig off the branch*	`[np v np pp(off)]`
		The twig broke off the branch	`[np v pp(off)]`
build	*carve*, *cut*	*Martha carves toys*	`[np v np]`
		Martha carves	`[np v]`
		Martha carved the baby a toy	`[np v np np]`

Table 5.1: Dorr and Jones' syntactic characterization of Levin classes

each of the Levin classes. The synonyms of the new verb were found within WordNet. The LDOCE grammar codes were found for both the new verb, and all the Levin classes that the synonyms belonged to. The Levin class was selected that has the closest match of LDOCE grammar codes to those of the new verb. In the event that there were no synonyms in WordNet that were also in Levin's classification, or there was a mismatch between the LDOCE codes for the new verb and those for the Levin classes, then Levin's system was simply augmented with a new class.

These experiments are of interest because they show the overlap between the syntactic aspects of Levin's classification and the semantic nature of her classes. They demonstrate how to extend this classification, with new classes for accuracy (Dang et al., 1998), and with new members and classes to increase coverage (Dorr & Jones, 1996). Crucially, extending the classification with new members using the approach of Dorr & Jones is dependant on manmade MRTs and MRDs. What is needed is a way of classifying new members when the information is not provided in a manmade resource. Using corpora by-passes reliance on the availability and adequacy of MRDs. The research of Resnik (1993a), Stevenson & Merlo (1999), Lapata (1999) and Schulte im Walde (1998) involved the classification of individual verbs according to corpus evidence.

Resnik (1993a) looked at the broad category of alternations in which direct objects are omitted, these are referred to by Resnik as implicit object alternations. He demonstrated a relationship between his measure of selectional preference, given in section 2.12 of this thesis on page 27, and the behaviour of verbs with regard to this alternation. Candidate verbs were selected which could occur with a direct object. These were manually classified as participants or not with help from the Collins COBUILD English Language Dictionary (Sinclair, 1987). Resnik then demonstrated a significant relationship between the selectional preference of a verb and its participation in the implicit object alternation for 34 test verbs. This finding supports the theoretical view that there is a connection between the capacity to omit an object and the ease with which the object's properties are inferred.

Resnik went on to show a correlation between a verb's selectional preference and the frequency with which it omits its objects. To do this he manually analysed 100 occurrences of 33 of the test verbs.[4] From this analysis, he determined the frequency of the implicit object construction. In the majority of cases, verbs with strong selectional properties had a higher number of occurrences of

[4]He excluded *have* for this experiment.

the implicit object construction than verbs with weaker selectional preferences. Although there were cases where verbs with strong selectional preference did not omit their objects, there were no counter examples of verbs with weak selectional preference omitting their objects frequently. This experiment showed that the ease of inference of the omitted object is a necessary condition for the construction.

In a third experiment, Resnik attempted a sub-classification of the object drop phenomena using his measure of selectional preference. The results were only significant for one out of two experiments, and so the evidence was not conclusive.

Stevenson & Merlo (1999) used syntactic and lexical cues for identifying verbal participation of 60 verbs in three verb classes, 20 verbs in each class. The verb classes were the unergative verbs, unaccusative verbs and 'object drop' verbs (i.e. those that take the implicit object construction). These three classes were chosen because they all involved a change in transitivity and a few well defined features can distinguish the three groups. The three classes are illustrated in the following examples:
Unergative:

(21) a. The boat floated over the lake.

b. The girl floated the boat over the lake.

Unaccusative:

(22) a. The sugar dissolved in the liquid.

b. The cook dissolved the sugar in the liquid.

Object Drop:

(23) a. The boy ate the food.

b. The boy ate.

Unergatives, like *float* are action verbs whose transitive form is causative, the action is caused by the subject and the subject of the intransitive becomes the object of the intransitive (Levin & Rappaport Hovav, 1995). Unaccusatives, like *dissolve* are change of state verbs where the transitive form is causative. Object drop verbs, on the other hand, do not have a causative form. The object is simply optional and is not present in the intransitive alternate.

Stevenson & Merlo used four linguistically motivated features to distinguish these groups. For example, unergative verbs are reported to be rare in the transitive form. These features were identified in the corpus using automatic POS tagging and parsing of the data. The features were:

1. VBD - main verb vs past participle
2. INTR - transitive vs intransitive use
3. ACT - active vs passive
4. CAUS - causative vs non-causative

The 60 verbs were manually selected from Levin's classification according to two criteria. Firstly, they were chosen by virtue of having sufficient frequency in a combined corpus (from the Brown and the WSJ) of 65 million words. Secondly, only verbs having one predominant intended sense in the corpus were used. Scores for the four features were obtained for the 60 verbs in the corpus. Relative frequency counts were used as scores for the VBD, INTR and ACT features. The frequency for the VBD feature was obtained for each verb using the relevant POS tag. The INTR feature was detected by searching for a nominal group after the main verb token. The ACT frequency score was obtained using the POS tag and the preceding auxiliary: *be* and the past participle tag signified a passive token. The CAUS score was calculated using the lemmas at the subject and direct object slots of the WSJ parses. The score was a ratio between the number of lemmas which occurred at both slots (duplicates included) divided by the number of lemmas in the two slots put together. The counts were normalised so as to give a score on a scale of 1 to 100 for each feature. The data for half of the verbs in each class was subject to manual scrutiny, after initial automatic processing. The rest of the data was produced fully automatically. The relative frequencies for the four features were then used to classify the four verbs automatically. A label was manually attached to each cluster using the class of the majority of verbs in the automatic cluster. The accuracy of automatic classification was 52% using all four features, compared to a baseline of 33%. If only the VBD, INTR and CAUS features were used the accuracy increased to 66%. All other combinations of features which were tried achieved accuracy levels between 45 and 54%. It would be interesting to know the difference in performance between the verbs for which human intervention was allowed, and those where the process was fully automated.

Stevenson & Merlo also used a supervised algorithm to further investigate the effect of the feature combinations. A supervised algorithm was used to avoid problems associated with manual labelling of the clusters. They obtained decision trees from the supervised data sets. In this case, the best results were obtained using all four features. It was possible to see the reduction in accuracy associated with removing each of the features in turn. Features that strongly correlated with one another were shown to affect accuracy least.

Lapata's research (1999) used both syntactic and semantic information to identify participation in diathesis. She aimed to investigate the extent to which alternations are attested in corpus data. She experimented with the dative (see example 15 above) and benefactive alternations (see example 24 below) using the entire BNC; 100 million words of written and spoken text. Lapata's strategy was to identify participants using a shallow parser and various linguistic and semantic cues. PP attachments were resolved using the lexical association score proposed by Hindle & Rooth (1993). Compound nouns, which could be mistaken for the double object construction, were filtered using the log-likelihood ratio test. The semantic cues were obtained by manual analysis. Lapata found that 81.5% of a sample of benefactive PPs could be categorized as animate, collective or denoted locations. She identified the corresponding WordNet classes and used these to identify benefactive PPs. The relative frequency of a frame for a verb, compared to the total frequency of the verb, was used for filtering out erroneous frames.

(24) a. She left a note for Duncan.

b. She left Duncan a note.

Recall and precision figures against a gold standard were not given for identification of participation. The emphasis was on the phenomena actually evident in the corpus data. From the corpus data, many of the verbs listed in Levin as taking an alternation were not found to have this alternation. This amounted to 44% of the verbs for the benefactive, and 52% for the dative. These figures only took into account the verbs for which at least one of the SCFs was observed. 54% of the verbs listed for the dative and benefactive by Levin were not acquired with either of the target SCFs. Conversely, many verbs not listed in Levin were detected as having an alternation using Lapata's criteria. Manual analysis of those verbs that are not in Levin revealed 18 false positives out of 52 candidates.

Lapata went on to use the relative frame frequencies to provide an estimate of the productivity of an alternation, for a semantic class, using a ratio based on that given in equation 5.1 devised by Briscoe & Copestake (1996). Briscoe & Copestake proposed dividing the number of verb types having entries in a lexicon matching the RHS of a lexical rule by the number matching the LHS. This was intended to indicate the ratio between the verbs that do participate and those that are potential candidates. Instead, Lapata used Levin's classification to define the possible candidates for an alternation. Productivity was calculated as in equation 5.7. The calculation is with respect to an alternation (A) and for the verbs in a specified Levin class ($classX$). This is the ratio of verbs from the Levin class automatically identified as participating, divided by the number of verbs in the class regardless of participation.

$$productivity(A|classX) = \frac{|verbs \in classX \cap verbs\ identified\ with\ A|}{|verbs \in classX|} \tag{5.7}$$

Lapata also used the frequency data to quantify the 'typicality' of an alternation, for a specified verb or verb class. The typicality measured the bias of the verb towards either of the frames involved in the alternation. Where there was no particular bias, the alternation was said to be typical for the verb. Lapata described this as the conditional probability of one particular frame given the verb, see equation 5.8. Lapata noted that, in her data, only two frames were involved in the denominator. This was because there were only two frames in the alternations studied. Her typicality measure was a proportion between the frequency of one frame and the sum of the frequencies of all frames *involved in the alternation*. However, she did not say how to determine which frame should be used for the numerator. This is an important issue. It is not obvious which form is the base form or, indeed, if any form is a base form (Montemagni & Pirrelli, 1995).

$$p(frame_i|verb) = \frac{\text{freq}(frame_i, verb)}{\sum_{j=1}^{n} \text{freq}(frame_j, verb)} \tag{5.8}$$

Schulte im Walde (1998) used automatically induced SCFs and selectional preferences to classify verbs according to their alternation behaviour. She acquired the SCF information using the statistical parser of Carroll & Rooth (1998). From this she obtained maximum probability parses for 5.5 million sentences of the BNC. She experimented with 88 SCFs which occurred more than 2000 times in the parses. The lexical heads were included in the parses, and these were used to obtain selectional preferences in WordNet, in the manner proposed by Ribas (1994, 1995b), which we outlined in section 2.3.1 in chapter 2. The preferences were collected at top level WordNet

classes. These classes were selected manually to give a high level of generalisation. Schulte im Walde then performed two different clustering algorithms on verbs with respect to the SCF data, each both with and without the selectional preference information. The first clustering algorithm was an iterative one, and relative entropy was used to compare the SCF distributions. The second method was the 'latent-class' EM-based algorithm proposed in Rooth (1998), and used in Rooth et al. (1999), which we described in section 2.2.2 of chapter 2.

Schulte im Walde evaluated the automatically induced clusters in terms of their agreement with Levin's classification. Recall was the percentage of verbs correctly assigned to a cluster which was a subset of the appropriate Levin class, compared to the total number of verbs which were clustered (153). Precision was the percentage of verbs which appeared in the correct cluster, compared to the number of verbs which appeared in any cluster. The iterative method achieved better results: 61% precision and 36% recall compared to 54% precision and 38% recall for the latent-class method. The latent-class method was, however, able to filter out multiple senses, whilst the iterative algorithm could only deal with verb types. These results were obtained using only SCF information. Surprisingly, adding the selectional preference information decreased precision and recall with both clustering algorithms. For example, for the iterative clustering, recall dropped from 36% to 20% with the selectional preference information, and precision fell from 61% to 38%. Schulte im Walde suggested that some work should be done to improve the choice of the conceptual classes in which the selectional preferences are represented. This should be done so that the classes representing the selectional preferences are more representative of the tokens occurring in the corpus data.

The work by Resnik, Stevenson & Merlo, Lapata and Schulte im Walde that we have just been considering all relates to the identification of verbs participating in diathesis alternations. Resnik's approach made use of a theory underpinning the implicit object construction. That verbs are licenced to omit their objects where the properties of those objects can easily be inferred. This is a useful means of identifying verbs which take the implicit object alternation. It cannot be transferred for use in identifying RSAs, however, it would be extremely useful alongside techniques for identifying other alternations.

The approach of Stevenson & Merlo hinges on the identification of linguistically motivated features that are relevant for the verb classification required. Such features were identified for the unergative / unaccusative / object drop distinctions. These features can be automatically detected in corpus data. Manual intervention was performed on the data for half the test verbs. The portability of this approach to new verb classes is dependent on the selection of the appropriate linguistic knowledge. Nevertheless, linguistically motivated features which act as salient cues for participation might easily be combined with methods requiring less a priori knowledge.

Lapata used both semantic and syntactic information to identify verbal participation in the dative and benefactive alternations. She, like Stevenson & Merlo, used linguistic heuristics for automatically identifying the alternations in corpus data. She went on to use her data for estimating the productivity of an alternation (using the corpus evidence for verbs in Levin's classification) and the typicality (using the relative frequency of the alternating frames).

Our approach resembles that of Lapata, except that we use automatically acquired selectional preferences instead of handcrafted semantic cues. Like Stevenson & Merlo, we look at overlap of

the lexical fillers of alternating slots. However, we show that this is useful for more than just the causative distinction and we do this using class-based preferences to avoid problems of sparse data. Like Schulte im Walde, we use selectional preferences automatically acquired within WordNet. However, her automatically acquired preferences degraded the performance of her system. We demonstrate that automatically acquired preferences are useful for detecting role switches. Like Lapata, Stevenson & Merlo and Schulte im Walde, we use automatically detected syntactic cues. In our approach, these are provided by the SCF acquisition system.

5.5 Combining Automatically Acquired Syntactic and Semantic Evidence for Diathesis Identification.

Diathesis alternations lie at the border between syntax and semantics. They are concerned with syntactic realizations that arise from the semantics underpinning the verb argument structure. We propose an automatic method that can be used for determining participation. We concern ourselves with alternations that are characterised by argument movement, whether or not this is accompanied by an omitted argument.

Selectional preference has already been linked with the capacity for verbs to drop their objects (Resnik, 1993a). This provided empirical evidence for the link between the ease at which the direct object can be inferred, and participation in the implicit object construction. The rationale underlying alternations concerned with argument movement are not so clear cut. For example, the conative alternation is shown in example 25. Below this are Levin's comments on this alternation:

(25) a. I pushed the table.

b. I pushed at the table.

> The conative alternation is a transitivity alternation in which the object of the verb in the intransitive variant turns up in the intransitive conative variant as the object of the preposition in a prepositional phrase headed by the preposition *at* (sometimes *on* with certain verbs of ingesting and the *push/pull* verbs). The use of the verb in the intransitive variant describes an 'attempted' action without specifying whether the action was actually carried out. The conative alternation seems to be found with verbs whose meaning includes notions of both contact and motion. (Levin, 1993, p.42).

Diathesis alternations arise from subtle semantic components of the verb [pp.4-11](Levin, 1993). However, these nuances of meaning are often rather elusive, and hard to define. They relate to the meaning of the verb, rather than the meaning of the arguments (Nicholls, 1995), although the combination is important for the alternation to occur (Montemagni et al., 1995). Identifying participation by searching for these semantic components automatically would be difficult for two reasons: (i) the semantic components are poorly defined, and (ii) there is no obvious way of labelling the verbs in the data without access to the knowledge that we are trying to discover.

In this chapter, we investigate whether diathesis alternations can be observed in corpus data by looking at role switching. That is, by seeing which verbs take arguments that can occur in slots with different grammatical roles in different syntactic realizations. We refer to these slots with different grammatical roles in the alternating variants as the target slots. For example, the target

slots of the causative alternation are the direct object slot of the transitive SCF and the subject slot of the intransitive SCF. In our experiments, we narrowed the search for potential candidates by using information from the SCF acquisition machinery. Verbs without the frames involved in a specified alternation were filtered out. Verbs where the frames occurred, but did so with a low frequency, were also filtered out. Identifying participation was then a matter of observing whether role switching took place between the target slots. There are however two main difficulties with this approach.

Firstly, the set of possible lexical fillers of one slot may not be the same as the set of fillers of another, even when the verb does participate. Montemagni et al. (1995) studied the causative-inchoative alternation in Italian. They found non-alternating argument heads in one of the target slots for many verbs, and in some cases in both. Interestingly, the non-alternating arguments were frequently related to the alternating arguments by figurative sense extensions. For example, alternating arguments of *suonare* (to ring or play) included *campana* (bell) and *musica* (music). Non-alternating arguments included *campanile* (clock tower) and *telefono* (telephone), which occurred as subjects only, and *Mozart* and *Dire Straits*, which occurred as objects. *Campanile* and *telefono* were seen as sense extensions of *campana* ('container for contents'). Meanwhile, *Mozart* and *Dire Straits* were seen as sense extensions of *musica* ('artist for art form'). In all cases, there was necessarily a non-empty intersection of the possible fillers of the alternating slots. It is our contention that automatic identification using information from the argument heads will be possible provided this intersection is sufficiently large.

The second difficulty for identifying RSAs using argument head data is due to sparse data. Many of the actual argument heads in the alternating slots may not overlap in the corpus data under scrutiny. To allow for this, we used the semantic preferences of the slots instead of the argument heads themselves. The semantic preferences provided generalisations of the type of argument head that can occur in a given slot. These preferences were automatically acquired as TCMs for the target slots of a RSA. We then compared the TCMs at the target slots. There were two ways in which we did this.

The first method did not use the TCMs themselves, but the calculations used for producing them. Using MDL for acquiring preferences provided us with a total cost, or description length, for each of the preference models. We exploited this in an approach which compared these costs for alternation detection. We refer to this as the 'MDL method' and describe it in section 5.5.2.

In the second method, we compared the TCMs at the target slots directly. We have three types of preference models (ATCM, PTCM and LLRTCM) available. All types provide a set of classes with a preference score at each class. Probability distributions are particular amenable for comparison using measures of distributional similarity. There are many established methods for comparing probability distributions. The details of some of these are provided in subsection 5.5.3 below. Sets of association scores and LLR scores could be normalised to produce a set of scores on a particular scale. However, since these scores are less well understood than probabilities, similarity measures were used only on the PTCMs. We refer to approaches for diathesis detection using distributional similarity measures as the 'similarity methods'.

We also investigated a method which used the argument heads directly. The method relied on a measure of overlap of the sets of lemmas at the target slots. We refer to this as the 'lemma based

method'. This can be viewed as a baseline method for comparison with the MDL and similarity approaches, which used the selectional preferences. This allows us to demonstrate the gains we make by generalising from lemmas to selectional preferences.

A lemma-based method will be particularly vulnerable to sparse data, since no generalisation is made. Additionally, because of the lack of generalisation, non-alternating arguments may be more problematic for a lemma-based approach than for a class-based approach. According to Montemagni et al. (1995), the lexemes that do not alternate are typically some figurative extension of ones that do. The figurative extensions, unlike the alternating arguments, do not typically form a semantically cohesive class. If this is the case, then non-alternating arguments should not pose too much of a problem for a class-based approach. The non-alternating arguments will be dispersed throughout WordNet. They will provide a certain amount of noise for selectional preference acquisition, but will not give strong preferences on the TCMs which might interfere with alternation detection.

5.5.1 Syntactic Information

We used syntactic information before we used the selectional preference models. We used the SCF lexicon for syntactic screening of verbs as candidates for a given alternation. To do this, we required a mapping between Levin alternations and the SCF classification used by our SCF acquisition system (Briscoe & Carroll, 1997). Such a mapping has been produced in draft form. [5] Levin alternations are listed alongside the SCF identifiers from Briscoe & Carroll's system. We hereafter refer to this mapping as the Levin–SCF mapping. This mapping provides several possibilities for alternating SCFs given a particular alternation. We use the frames most prototypical of the Levin alternations. For example, the most prototypical, mapping for the conative alternation (given above in example 25) is given as: [6]

(class 87_96)	↔ (classes 24, 24_50 or 24_51_161)
`[np1 v pp2(at,on)]`	↔ `[ np1 v np2]`

Another possibility is also suggested, with an additional PP argument:

(class 77_95)	↔ (classes 56_49, 49_50 or 31_49)
`[np1 v pp2(at,on) pp3]`	↔ `[np1 v np2 pp3]`

This second possibility was presumably added in case a PP has been attached to the verb by the parser, perhaps in error, and the SCF acquisition system has not filtered this out as an adjunct. We do not include these cases. They appear to have been included to try and compensate for errors made by the parser and SCF acquisition system.

[5] The mapping was the work of Anna Korhonen. We are indebted to her for the use of this.

[6] Each distinct SCF classification is represented by one or more SCF class numbers joined together with an underscore e.g. 24_51_161. More than one class is provided by the SCF acquisition system where, for some subcategorization patterns, the system cannot tell which of the classes is appropriate so the possible classes are conjoined. Furthermore the mapping specifies one or more of these classifications where any of them would be appropriate for this alternation. For example, The transitive SCF is identified by any of the three specified classes, 24, 24_50 or 24_51_161.

5.5.2 Using MDL for Diathesis Detection

This section outlines a method for diathesis detection which uses the cost of producing the preference models, rather than using the actual preference models themselves. A comparison is made between the sum of the costs for separate TCMs at the target slots, and the cost of a TCM for the combined data from both the target slots (the combined model). If the data at the two slots is similar, then the cost of the combined model is smaller than the sum of the costs for the separate models. This method assumes an implicit threshold at the cost of the two separate models. Participation is detected in cases where the combined model is cheaper than the separate models.

In our preference acquisition system, MDL selects the model which makes the best compromise between the detail of the model, and the match of the model to the data. The cost of the model is calculated using the model description length. The cost of the data encoded in the model is calculated using the data description length. The sum of these two description lengths (the total description length) is minimised. As we explained in section 3.6.4, the data description length, and therefore the total description length, is reduced in cases where the nouns senses in the corpus data accumulate in the same classes in WordNet. The more homogeneous the data is, the cheaper it becomes to store it. Diathesis alternation detection works by using the costs to indicate where the data across two slots is similar, since combining it is cheaper than modelling it separately.

For the three different model types (PTCMs, ATCMs and LLRTCMs), the description lengths are calculated differently. They do however share the characteristic of cheaper costs for more homogeneous data. For the PTCMs this is because the description length (given in equation 2.17 on page 33) is obtained using the log of the conditional probability ($p(c|v)$). When the probability distribution becomes uneven the message becomes more predictable. The entropy decreases and encoding the message becomes cheaper.

The ATCM description length (given in equation 2.20 on page 34) is based on the description length for the probabilistic models. It is also reduced where the verb specific data is concentrated in some areas. This is because classes with a high conditional probability, compared to the prior probability, have a high association score. High association scores contribute to a low description length.

The LLRTCM description lengths (given in equation 2.21 on page 38) are also reduced by high concentrations of the conditional probability distribution at particular classes provided that the observed probability distribution exceeds that expected for these classes. The expected value is based on the null hypothesis that the verb does not have an effect on the probability distribution. High concentrations of conditional probability tend to increase LLR and decrease the relative cost of these models.

Figure 5.1 illustrates the scenario of diathesis alternation identification using ATCMs. Three ATCMs were acquired for each predicate. The figure displays the costs of the three models actually used for detection of the causative alternation for *begin*. That is the minimum costs found when using equation 2.20 on page 34 to select the optimum model for the three respective datasets. One model was acquired for the data from the object of the transitive frame. [7] The cost of this model was -552. Another model was acquired for the subject of the intransitive. The cost of this model was -824. Finally, the third ATCM was acquired for the combined data from the two frames. The

[7] We collected data from both active and passive versions of this frame.

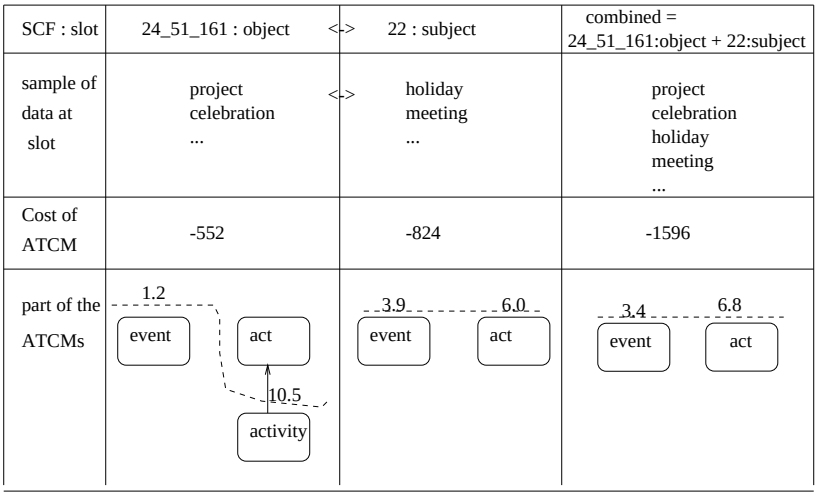

Figure 5.1: Causative detection for the verb *begin*

cost of this combined model was -1596. ATCMs are produced as a by-product of the procedure for estimating a model for the conditional data (Abe & Li, 1996). The use of $-\log A(c,v)$ (see equation 2.20 on page 34) in the data description length frequently results in a negative cost. The costs of the separate models can nevertheless be added together to give the cost of encoding the data separately. In the case of *begin*, we have $-824 + -552 = -1376$ when summing the costs at the target slots. This is a higher value than -1596, taking the sign into account, indicating that it is cheaper to combine the data into one model for *begin*. We therefore conclude that *begin* participates in the causative alternation.

The TCM type has a considerable effect on diathesis alternation detection, despite the fact that all model types share the characteristic of being cheaper for homogeneous data. This is because the description length calculation for ATCMs and LLRTCMs departs from a true MDL description length, measured as the number of bits required.

As we just saw, the ATCM cost can be negative. This is related to the number of bits required only indirectly when considered alongside the prior model (Abe & Li, 1996). A greater cause for concern is that the ATCM depends heavily on the prior TCM that is used. [8] This is particularly important when using the costs for identifying alternations, because a decision needs to be made as to which slot to obtain the prior data from, for the combined model. The results reported in (McCarthy & Korhonen, 1998) were heavily dependent on the prior used, as we report below.

The LLRTCMs use a more discerning score, as far as low frequency data is concerned. However, the description length calculation combines LLR scores across the TCM. This is done as a heuristic; there is no theoretical justification for doing so. As a consequence, the costs do not reflect the actual description length required for encoding the model and data. Therefore, the costs of two separate models cannot legitimately be combined for comparison with the cost of the combined model.

The PTCMs have a clear description length and do not require a separate model for the prior distribution. Therefore, these are straightforward to use for diathesis detection.

[8] Hang Li, personal communication.

There are two foreseeable problems with the MDL approach, regardless of the TCM type, both of which can give rise to false positives. The first problem also applies to the similarity and lemma based methods. The semantic constraints at the target slots may be similar without this being due to an alternation. For example, the causative alternation involves the object of the transitive frame switching place with the subject of the intransitive (as in example 13 on page 110). Some verbs may have similar argument heads in the target slots, without participating. For example, *help* commonly occurs with nouns or pronouns under the **person** class in both the subject and object, as in example 26. The semantic roles of the subject of the intransitive (agent) and object of the transitive (theme) are different. However, the overlap between the semantic type of the filler of these slots might lead to false identification of participation. The extent of this problem depends on the extent of the similarity of the fillers in the different semantic roles involved. It is unlikely that there will be a total overlap of lexical fillers in two slots having different semantic roles. For example, although nominals denoting **person** frequently occurred as the subject and direct object of *help* in lexicon D, nominals denoting **sum of money** occurred as subjects but not as direct objects.

(26) a. I help him.

b. He helps.

The second cause of false positives is peculiar to the MDL method. The relative frequencies of the alternate variants can be substantially different. When this is the case, the argument head data from one frame can overwhelm the other. False positives arise because the cost of the combined model is close to the cost of the model for the overwhelming frame. This is a cheaper alternative to separate models which require separate model description lengths.

To alleviate the first source of false positives, we investigated whether filtering out candidates which have similar semantic constraints where the two slots co-occur in the same frame improved accuracy. This was only possible for alternations where the slots do co-occur in the same frame. An example where they do is the transitive variant of the causative alternation. For this alternation, the object of the transitive frame is predicted to have similar constraints to the subject of the intransitive. If there are similar semantic constraints in the two grammatical slots in the transitive frame to start with, then we cannot be sure whether a similarity at the slots in the two SCFs is an indication of alternation. The intransitive frame could instead be an instance of the implicit object construction. For this reason, we compared accuracy with and without filtering out candidates with similar subjects and objects in the transitive.

The second problem is a particular obstacle for the MDL method, because this is affected by the sample size. When there is a disparity in the relative frequencies of the alternating frames, the cost of the TCM at the more frequent slot is likely to overwhelm the cost at the other slot because of the larger data description length. This renders the MDL method for establishing participation vulnerable, except for predicates with an even ratio between the alternating frames. The similarity approaches are more suited for predicates with an uneven ratio.

5.5.3 Measuring Similarity between Semantic Preferences

Here we discuss some of the measures available for comparing probability distributions from the PTCMs for the similarity-based approaches. Most are measures of distance, rather than similarity,

but this does not affect their utility. What we require is a means of ranking the verbs on grounds of the similarity, or dissimilarity, of the selectional preferences in the target slots.

We describe below some of the similarity measures used in the literature. We did not smooth our selectional preference models prior to using these similarity measures. We did not do so because using a class-based approach was an alternative to smoothing (Resnik, 1993a; Li & Abe, 1995). The majority of PTCMs did not contain classes with zero probabilities. However, even using a class-based approach there were some classes with a zero conditional probability. We coped with these by only considering similarity measures which are defined for zero values. We investigated a variety of measures to see if there was a significant difference in performance depending on the measure used.

Lee (1997, 1999) has compared the performance of a variety of similarity measures, for the purpose of smoothing language models. She used a co-occurrence pair decision task, like the pseudo-disambiguation experiments in section 4.3.4 of chapter 4. The system had to identify which of two verbs was more likely to co-occur with a given noun. The correct decision was the one attested in the corpus. Her experiments showed that high performance is associated with similarity measures which concentrate effort on items for which both probability estimates under comparison are non zero (1999). In the following, we note the similarity measures which Lee reported to be most affected by zero values.

Euclidean Distance

Euclidean distance (ED) is a positive valued metric. It represents the geometric distance between two vectors. A large value represents a large distance (or dissimilarity). It is given in equation 5.9.

$$\text{ED}(p1(x), p2(x)) = \sqrt{\sum_x (p1(x) - p2(x))^2} \tag{5.9}$$

This function is affected by zero probabilities (Lee, 1999, 1997). Nevertheless, since this function is defined even when there are zero values we took it forward for experimentation.

Cosine

The cosine is related to the angle between two vectors. It is defined in equation 5.10. It is a true similarity measure, rather than a measure of disimilarity. Higher scores represent greater similarity. The range of possible values is between 0 and 1. The value is 1 where $p1(x) = p2(x)$ for all values of x. The minimum value is 0. This results where one vector has zero values for every value of x that the other vector has a non zero value. It is useful in cases where there are differences of scale (Schütze, 1992) although this was not the case here, since we were using probabilities.

$$cos(p1(x), p2(x)) = \frac{\sum_x p1(x)p2(x)}{\sqrt{\sum_x p1(x)^2}\sqrt{\sum_x p2(x)^2}} \tag{5.10}$$

Lee noted that this value, whilst placing most importance on values of x with non zero estimates for both probability distributions, does also use the items which have a non zero value for one of the estimators. We also took this measure forward for experimentation.

L_1 norm

This is a geometrically-motivated measure and is given in equation 5.11. It is also known as the 'Manhattan' or 'taxi-cab' distance (Lee, 1997).

$$L_1(p1(x), p2(x)) = \sum_x |p1(x) - p2(x)| \tag{5.11}$$

This function can be expressed in terms of only the values of x which have non zero values for both p1 and p2 (Lee, 1997, p13). This form is given in equation 5.12, where the term $\sum_{x \in p1p2}$ is a sum over those values of x which have a non zero probability for both $p1$ and $p2$. This function is not greatly affected by zero estimates. The L_1 norm is taken forward for our diathesis detection experiments.

$$L_1(p1(x), p2(x)) = 2 + \sum_{x \in p1p2} (|p1(x) - p2(x)| - p1(x) - p2(x)) \tag{5.12}$$

Cross entropy

Cross-entropy is often used in the evaluation of language models. It is defined in equation 5.13. It is an indicator of how good one distribution ($p1(x)$) is as an approximation for another ($p2(x)$).

$$\text{cross entropy}(p1(x), p2(x)) = -\sum_x p1(x) \log p2(x) \tag{5.13}$$

Cross entropy is minimal when $p1(x) = p2(x)$. Thus, this is a measure of dissimilarity. This measure is not defined for zero values and was therefore not a candidate for diathesis alternation detection. It is included here as it bears a relation to relative entropy.

Relative entropy

This is also termed Kulback-Liebler distance (Cover & Thomas, 1991) and is given in equation 5.14. It measures the average cost of using one distribution to code for the other. It can also be defined in terms of cross-entropy (Krenn & Samuelsson, 1997; Manning & Schütze, 1999) as in equation 5.15

$$D(p1(x)||p2(x)) = \sum_x p1(x) \times \log \frac{p1(x)}{p2(x)} \tag{5.14}$$

$$D(p1(x)||p2(x)) = \text{cross entropy}(p1(x), p2(x)) - \text{entropy}(p1(x)) \tag{5.15}$$

$$\text{where entropy}(p1(x)) = -\sum_x p1(x) \log p1(x) \tag{5.16}$$

Unfortunately, this measure is undefined where there are non zero values for p1(x) which have corresponding zero values for p2(x). For this reason, we do not suggest it for diathesis alternation detection using non-smoothed models.

α-skew divergence

The α-skew divergence (αSD) was devised by Lee (1999) and is defined in equation 5.17. This measure is a modification of Kulback-Liebler divergence. The α constant is a value between 0 and 1 which smooths p1(x) with p2(x) so that the αSD is always defined. If α is set to 1 then this measure is equivalent to the Kulback-Liebler divergence.

Lee compared the performance of this measure with a variety of similarity measures. This was done on the co-occurrence pair decision task described above. α-skew divergence had a statistically significant error reduction compared to all the other similarity measures used.

Lee used a value of $\alpha = 0.99$. She suggested that the value of α selected should be inversely related to the sparseness of the data. This is yet to be proved. We took α-skew divergence forward for diathesis alternation detection, using the same value (0.99) for α as Lee.

$$\alpha\text{SD}(p1(x), p2(x)) = D(p2(x)||((\alpha \times p1(x)) + ((1-\alpha) \times p2(x)))) \tag{5.17}$$

5.5.4 The Lemma-Based Approach

We used a lemma based approach to compare with the MDL and similarity-based approaches. This was used to demonstrate that generalisations to classes provide superior results to those obtained when using the argument heads directly. We used a measure incorporating the proportion of overlap between the argument heads in the two target slots. The measure was termed lemma overlap (LO) and is given in equation 5.18, where A and B represent the sets of lemmas at the two target slots. LO is defined as the size of the intersection (duplicates included) of the multisets [9] of argument heads at the target slots divided by the size of the smaller of the two multisets. For example, in diagram 5.2 a Venn diagram illustrates the sets of the lemmas at the transitive and intransitive SCFs for the verb *break*. The intersection of two multisets includes duplicate items only as many times as the item is in both sets. For example, if one slot contained the argument heads {*person, person, person, child, man, spokeswoman*}, and the other slot contained {*person, person, child, chair, collection*}, then the intersection would be {*person, person, child*}, and LO would be $\frac{3}{5}$. This measure ranges between 0 (no overlap) and 1 (where one set is a proper subset of that at the other slot).

$$\text{LO(A, B)} = \frac{|multiset\ intersection(A\,B)|}{|smallest\ set(A,B)|} \tag{5.18}$$

Our measure bears some relation to the measure used by Stevenson & Merlo for detecting the CAUS (causative) feature. They used the size of the 'overlap' of the two multisets, but where duplicated items were included as many times as there were duplicates in one of the sets. Using the above example, the overlap between {*person, person, person, child, man, spokeswoman*} and {*person, person, child, chair, collection*} is {*person, person, person, child*} (4). Stevenson & Merlo divided this by the size of the union of the two multisets, (11, in this example). This gave a postive number at or above 0 and below 1. This proportion will be larger as the overlap increases, but this is tempered by the size of the union. The proportion will never reach 1 because the size of

[9] A mutliset is a set which may contain items more than once,

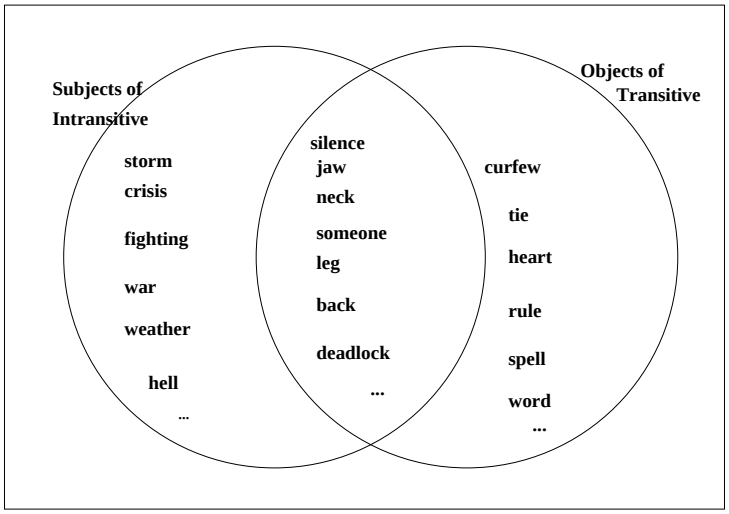

Figure 5.2: Lemma causative detection for the verb *break*

the union is always larger than the size of the overlap. This is because the overlap only includes duplicated items as many times as they are in *one of the multisets*. At least one duplicated item will always be counted in the union, but not the overlap. Moreover, if two sets are identical, then the score will only be 0.5. We counted duplicated items only as many times as they exist in both sets to give a larger score where there is more evidence of a lemma occurring in both slots. In the case that one multiset is a subset of the other, our LO score is 1.

5.5.5 Identification of Participation

Of the three methods, the MDL method was the only one where a threshold was not explicitly required to determine participation. There was an implicit threshold which was determined by the cost of the two separate models. Participation was predicted where the cost of the combined data was lower than this implicit threshold.

Accuracy was calculated for the MDL method by comparing the system's decision against the decision of human judges. The system decided in all cases presented to it.

For the similarity and lemma-based experiments, the scores were on a continuous scale. We used the Mann Whitney U test (Siegel & Castellan, 1988) to determine the significance of the relationship between the scores and participation.[10] Accuracy was calculated using a threshold at the mean, or median to determine participation. The decision for each verb was compared to the consensus of the human judges. The thresholds were used simply to give us a rough idea of how well the measures partitioned the verbs. If a threshold were to be used in earnest then it would need to be obtained from held out data.

5.6 Scope

Our strategy of using SCF and selectional preferences is applicable to a subset of the alternations described in Levin. In this section, we demarcate the boundaries for our approach. The alternations

[10]The Mann Whitney U test was selected since the scores of both the positive and negative verbs were significantly skewed, according to the measure of skewness given in Howitt & Cramer (1997, p340).

Alternation	Levin Section	Information
Middle	1.1.1	Adverbial
Reciprocal	2.5	co-ordinated nouns in NP
Body-Part Possessor Ascension	2.12	possessive
Possessor-Attribute Factoring	2.13	possessive
Reflexive	4	reflexive
Postverbal subjects	6	There + subject analysis

Table 5.2: Alternations requiring additional syntactic information

which do not meet our criteria are identified and eliminated. We also point out others, which are not considered here, but which could be identified with some modifications to the SCF acquisition process. We refer to the alternations using Levin's (1993) terminology, and we supply the section number of the classification from her book in brackets. For example the causative-inchoative alternation is listed in the book as (Levin, 1.1.2.1).

In our experiments, we cannot distinguish subtypes of alternations which are not reflected in different SCFs. So, for example, we experiment with the causative alternations collectively, rather than using the finer distinctions made by Levin.

The alternations we are concerned with involve role switches. We detect these using selectional preferences at the target slots. Alternations concerned solely with arguments which are omitted are not handled here. These are Levin's 'Unexpressed Object Alternations' (Levin 1.2). We referred to them above collectively as the implicit object alternation when we discussed Resnik's (1993a) work.

We did not include alternations which require syntactic information that is not currently stored in the SCF lexicon. For this reason, we did not include alternations which typically involve an adjunct, since the SCF filters these out. The middle alternation was therefore ignored (Levin 1.1). We also did not look at alternations which involve analysis at the phrase level, for example, the 'reciprocal' alternations (Levin 2.5). An example of this is given below in example 27, which is an instance of the simple reciprocal alternation (transitive) (Levin 2.5.1). This alternation includes a collective NP, *the sugar and the butter* in our example. Identifying information at the phrase level is lost in the process of SCF acquisition. This information could be recovered from the parser. Whilst we did not experiment with these alternations, our method is still applicable. The syntactic markers are in many cases easy to identify. For example, Body-Part Possessor recognition would be possible by including the possessive marker in our enumeration of argument heads. Where the syntactic markers can be identified, candidates which do not participate will more readily be filtered out. In table 5.2, we list alternations which would require us to store additional syntactic markers in our argument head entries. The final column indicates the syntactic information that is require for correct identification of these alternations.

(27) a. I mixed the sugar into the butter.

b. I mixed the sugar and the butter.

We did not experiment with the passive (Levin 5). The passive is nearly exceptionless in

Parser	Zero crossings (% sents.)	Mean crossings per sent.	Bracket recall (%)	Bracket precision (%)
LR	57.2	1.11	82.54	83.00
PCP	54.2	1.13	82.50	82.68

Table 5.3: Parser evaluation

its application to transitive verbs. Transitive verbs are identified as such by the SCF acquisition system. Also, the parser already identifies passives using the past participle and the auxiliary *be*.

There are a number of miscellaneous constructions listed by Levin (Levin 7). These concern the semantic type of the arguments. These might best be dealt with by looking at the actual semantics of the argument heads. Also, a further section (Levin 8) deals with special diatheses, not featuring in alternations. These are not handled here.

5.7 Sparse Data Problems

Our methods are applicable to all the alternations listed in Levin, but not mentioned in section 5.6 above. However, we were not able to experiment with all of these because of sparse data problems. We experimented with two lexicons: Lexicon A and Lexicon D. Lexicon A is smaller than Lexicon D. It was built using 10.8 million words of parsed text from the BNC. Lexicon D was constructed using 19.3 million words of parsed text from the same source.

The parses used for lexicon A were obtained from a probabilistic chart parser (PCP) (Chitrao & Grishman, 1990). The parses for lexicon D were the output from a LR parser (Inui et al., 1997). The parsers were evaluated on a test suite of 500 manually bracketed sentences. [11] The table 5.3 displays the results of this evaluation. The column labelled 'zero crossings' gives the percentage of sentences for which no bracketing produced by the parser overlaps with any in the gold standard test suite. 'Mean crossings' indicates the mean number, per sentence, of brackets from the parser which overlap with the gold standard brackets. 'Bracket recall' shows the number of times the system's brackets matched those of the gold standard, divided by the number of brackets in the gold standard. 'Bracket precision' shows the number of times the system's brackets matched those of the gold standard, divided by the number of brackets in the parser output. The LR parser is slightly more accurate than the PCP, but the differences are not statistically significant.

To investigate if our methods worked, we needed a sample of positive and negative candidates each with sufficient data for preference acquisition.[12] For experimentation, we needed to be sure that our method successfully separated the positive and negative candidates for each alternation. To be sure of this, we applied significance tests to see if there was a significant relationship between our similarity and lemma overlap measures and participation. We used an even number of positive and negative candidates which gave us a 50% baseline for a random classification. We limited experimentation to alternations for which we had at least 3 positive and 3 negative candidates. This just met the minimum requirement for group size for significance testing with the Mann

[11] We are indebted to John Carroll for the parses and evaluation results.

[12] We used a threshold of 10 argument heads which can be classified in WordNet.

Whitney U test.

Table 5.4 lists the alternations with which we did not experiment because of sparse data. The sparse data problem was made worse by the fact that some alternations only exist for a very small number of verbs. In the worst case, experimentation was impossible for individual alternations because there was only one candidate. This is the situation for the blame alternation (Levin 2.10) given in sentence 28. This alternation only holds for *blame*.

(28) a. Mira blamed the accident on Terry.

b. Mira blamed Terry for the accident.

Many other alternations are listed in Levin for a small set of low frequency verbs. For example, the substance/source alternation (Levin 1.1.3) exemplified in 29.

(29) a. Heat radiates from the sun.

b. The sun radiates heat.

Levin lists participant as :
belch (12), *bleed* (82), *bubble* (64), *dribble* (12), *drip* (73), *drool* (9), *emanate* (64), *exude* (30), *gush* (30), *leak* (84), *ooze* (41), *pour* (449), *puff* (30), *radiate* (55), *seep* (65), *shed* (125), *spew* (8), *spout* (5), *sprout* (43), *spurt* (14), *squirt*(4), *steam* (64), *stream* (64), *sweat* (67)

The frequencies for these verbs from Lexicon A are given in brackets. These are frequencies over all SCFs. Once we are specific to the SCF, the data is even more sparse. The most frequent verb was *pour*. The [`np v np`] and [`np v pp(from)`] SCFs involved in the substance/source alternation are specified by the Levin–SCF mapping. The frequencies for *pour* were 138 ([`np v np`]) and 14 ([`np v pp(from)`]). For most verbs, some argument heads were proper nouns, which we did not use, or common nouns unclassifiable in WordNet. After *pour*, the next most frequent candidate, *shed* was not identified as taking the PP SCF in the lexicon. From this example, it is easy to see how sparse data hampers alternation detection. Particularly where alternations involve specific prepositions.

Additionally, we determined the grouping for evaluation using the decisions of human experts. This is described below in section 5.8.2. If there was a significant level of disagreement between judges, then the alternation was not used. Furthermore, if there was strong disagreement from the human judges for a particular verb then it was not used. Thus, many alternations were not investigated because there were not enough suitable candidates in the corpus data available.

The alternations with which we experimented are listed in table 5.5

5.8 Diathesis Identification Experiments

In this section, we present our diathesis identification experiments. In subsection 5.8.1 we say a little more about the selection of candidates using syntactic information. Subsection 5.8.2 describes how the gold standard was set up for evaluation. Before the sections containing the main bulk of our results, we provide a subsection (5.8.3) which discusses the alternations which were identified using only SCF information. The selectional preferences were not required for these alternations.

Alternation	Levin Sect. No
Substance/Source	1.1.3
Locative prep drop	1.4.1
With prep drop	1.4.2
Locative Alternations	2.3
Creation and Transformation	2.4
Fulfilling	2.6
Image Impression	2.7
With/Against	2.8
Through/With	2.9
Blame	2.10
Search	2.11
As	2.14
Oblique subject	3

Table 5.4: Levin alternations with sparse data

Alternation	Levin Sect. No
Causative	1.1.2
Conative	1.3
Dative	2.1
Benefactive	2.2

Table 5.5: Candidates for experimentation

The main results are divided into four subsections. Each relating to one of the three approaches outlined in section 5.5, with two of the subsections corresponding to the MDL approach. Results from the lemma method are given in subsection 5.8.4. Those from the MDL method using the ATCMs are given in 5.8.5, whilst those using the PTCMs are in 5.8.6. Finally, the results from the similarity methods are in 5.8.7.

5.8.1 Using the Syntactic Information

For experimentation, we required alternations with positive and negative candidates. This was necessary to investigate the success of our methods. The SCFs required for each alternation were specified by the Levin–SCF mapping. Candidates were selected which had the appropriate SCFs, and where each SCF was listed with 10 or more argument heads which could be classified in WordNet. In some cases, MDL recommended a TCM at the dummy root. For the PTCMs, we used the WordNet root classes, below the root, in these cases.

For alternations involving PPs, the alternation is sometimes only valid for specified prepositions. For example, the conative alternation is only relevant for the prepositions *on* and *at*. Moreover, a verb might undergo the alternation with some of the prepositions specified for an alternation, but not all. For alternations involving prepositions we needed to specify the preposition, as well as the verb, that the selectional preferences were collected for.

For the lemma and MDL experiments we used lexicon A. For the similarity approach we used lexicon D.

5.8.2 Human Agreement

Candidates were selected for the alternations in table 5.5 by virtue of the SCFs they were credited with having in the lexicon. We made an initial decision on participation for each verb with the appropriate frames for a given alternation. This was to ensure that we had an even split between positive and negative candidates in the test set which we presented blind to our human judges. The test set was chosen from these positive and negative groups, with the same number of candidates taken at random from each group. The entire test set was presented to the judges as an alphabetically sorted list. The judges had to stipulate whether each candidate participated in the specified alternation or not. A 'do not know' verdict was permitted.

For the MDL and lemma-based experiments, two human judges were used. Candidates were selected where these judges were in total agreement. Verbs for which the judges disagreed were removed.

For the similarity approach, the decisions of four judges were obtained. The kappa statistic (Siegel & Castellan, 1988) was calculated to establish whether there was significant agreement between judges. This statistic was calculated using the number of judges assigning each category (positive, negative or don't know) to each verb. In our equations, n_{ij} represents the number of times category j was assigned to the verb i. The kappa statistic (K) is the ratio of the proportion of times that the judges agreed (corrected for chance agreement) to the proportion of times that the judges could have potentially agreed (again, corrected for chance agreement). This ratio is given in equation 5.19.

$$K = \frac{P(A) - P(E)}{1 - P(E)} \tag{5.19}$$

Chance agreement, $P(E)$, is defined in equation 5.20, where m is the number of categories (3 in our case) , k is the number of judges and N is the number of verbs.

$$P(E) = \sum_{j=1}^{m} (\frac{\sum_{i=1}^{N} n_{ij}}{Nk})^2 \tag{5.20}$$

The proportion of times that the judges agreed $(P(A))$ is defined in equation 5.21.

$$P(A) = \frac{1}{Nk(k-1)} \sum_{i=1}^{N} \sum_{j=1}^{m} n_{ij}^2 - \frac{1}{k-1} \tag{5.21}$$

K ranges in value between 0 (no agreement) and 1 (total agreement). The value of the kappa statistic was used for significance testing to see if the judges agreement was greater than or less than that expected by chance. For large values of N, K is normally distributed with variance as in equation 5.22:

$$var(K) \approx \frac{2}{Nk(k-1)} \frac{P(E) - (2k-3)[P(E)]^2 + 2(k-2)\sum_j (\frac{\sum_{i=1}^{N} n_{ij}}{Nk})^3}{[1 - P(E)]^2} \tag{5.22}$$

Equation 5.23 gives the value to be looked up in the tables for the normal distribution to determine the probability of the observed agreement being due to chance.

$$z = \frac{K}{\sqrt{var(K)}} \tag{5.23}$$

We did not attempt automatic identification of participation where there was not significant agreement between the human judges. For alternations for which there was a significant level of agreement, only verbs with 75% agreement or more (3 judges or more) were taken forward for the diathesis alternation identification experiments.

5.8.3 Alternations Identified Using Only Syntactic Information

From experiments with the data in Lexicon A, two of the alternations under investigation had only positive candidates with sufficient frequency at the alternating SCFs (McCarthy & Korhonen, 1998). These alternations were the dative and benefactive alternations. The positive candidates were acknowledged as such by the human judges. Participation in these alternations was therefore detected by syntactic information alone. Interestingly, these were the alternations which Lapata (Lapata, 1999) identified using shallow parses, statistics and a variety of linguistic cues. Lapata did this using a larger sample (the full BNC).

Dative

The potential candidates remaining with more than 10 instances at the two slots stipulated for the dative by our alternation classification were:

award, *give*, *hand*, *lend*, *offer*, *owe*.

These verbs all take the dative alternation. Thus, the SCF lexicon and Levin–SCF mapping alone was enough to predict alternation in these cases. There were other verbs in the corpus data that can take the dative according to Levin (Levin, 1993). However, they did not occur in the corpus with sufficient frequency for detection of both the SCFs.

Benefactive

The frames characterising the benefactive alternation were also enough to filter out irrelevant verbs. The verbs with sufficient occurrences of the SCFs were:

award, *earn*, *give*.

This information could be useful for further refining the SCF classification. The SCF stipulated by the Levin–SCF mapping for the prepositional construction of the dative and benefactive is also assigned by the SCF acquisition system to other `[np v np pp]` constructions. The co-occurrence of this SCF with the SCF for the double object construction could be used to provide specific SCF labels for the prepositional phrase construction of the dative (*to*) and benefactive (*for*). These could then be distinguished from more general `[ np v np pp]` constructions.

5.8.4 Lemma-Based Experiments

We investigated a lemma-based approach for diathesis alternation detection using the LO measure given above in equation 5.18 in section 5.5.4. We investigated the relationship between this value and participation in the causative and conative alternations. This was done using lexicon A.

Causative

For the causative alternation, there were plenty of candidates that occurred with the alternating frames. We took a sample of 54 positives and 56 negatives. These occurred in lexicon A with sufficient frequencies for the alternating SCFs to be used for the experiments requiring selectional preference acquisition. These were those with 100% agreement between two judges. The sample is given here:

Positive sample (54 verbs):-

bake, *begin*, *bend*, *blend*, *boil*, *break*, *burn*, *calm*, *change*, *clear*, *close*, *continue*, *cook*, *crack*, *crash*, *decrease*, *develop*, *drive*, *drop*, *dry*, *end*, *expand*, *finish*, *flood*, *fly*, *grow*, *hang*, *hurt*, *improve*, *increase*, *land*, *march*, *match*, *melt*, *mix*, *move*, *open*, *record*, *repeat*, *ring*, *roll*, *settle*, *shut*, *sink*, *split*, *spread*, *start*, *stop*, *stretch*, *swing*, *train*, *turn*, *vary*, *wake*.

Negative sample (56 verbs):-

add, *admit*, *answer*, *ask*, *attack*, *believe*, *bother*, *catch*, *charge*, *choose*, *climb*, *compare*, *cost*, *cut*, *declare*, *demand*, *dress*, *drink*, *eat*, *expect*, *feed*, *feel*, *hear*, *help*, *hide*, *imagine*, *imply*, *investigate*,

Table 5.6: Mann Whitney U test results for conative

	on		*at*		*on & at*	
group	size	sum	size	sum	size	sum
positive	4	22	4	23	8	86
negative	4	14	4	13	8	50

kick, *know*, *like*, *live*, *love*, *miss*, *nod*, *observe*, *pack*, *pass*, *pay*, *perform*, *plan*, *pull*, *read*, *remain*, *remember*, *shout*, *sing*, *steal*, *survive*, *suspect*, *think*, *understand*, *wash*, *win*, *work*, *write*.

The Mann Whitney U Test was used on the LO scores. A one tailed test was used. The null hypothesis was that there is no relationship between LO and participation. The alternative hypothesis was that the LO scores for participating verbs are greater than those of non-participating verbs.

The Mann Whitney U test for large samples was used. From this a value of z was calculated and this was looked up in tables of the normal distribution. The z score obtained was 1.007. For the right tailed test a z score of this magnitude or more has a probability of 0.16 of occurring by chance if the null hypothesis is true. This was not a significant result. [13] There was not a strong enough relationship to reject the null hypothesis.

Conative

For the conative we performed two experiments. One for participation with the preposition *on* and one for *at*. Four positive and four negative candidates were chosen for each preposition. These were:

positive sample *on*:
bang, *cut*, *press*, *pull*.

negative sample *on*:
agree, *move*, *remain*, *work*.

positive sample *at*:
pull, *push*, *shoot*, *tug*.

negative sample *at*:
call, *move*, *remain*, *work*.

The sample size was rather small. We were limited by the number of positive candidates available with sufficient frequencies at the alternating SCFs. In addition to this, we removed candidates where our two human judges were in disagreement. The Mann Whitney U test for small samples was conducted.

[13] We would expect a probability below 0.05 for a significant result, if we carry out the test to the 95% level.

Conative on

The results of the Mann Whitney U test signified that again, there was not a significant relation between LO and participation. The group size and sum of the ranks that we obtained are given in table 5.6 in the two columns headed by *on*. The probability of getting the sum of the ranks for the positive group at 22 or above was 0.17. This was not significant.

Conative at

There was not a significant relationship between LO and participation, according to the Mann Whitney U test. The group size and sum of the ranks obtained from our experiment are given in table 5.6 in the two colums headed by *at*. The probability of getting the sum of the ranks for the positive group at 23 or above was 0.1. Again, this was not significant.

Conative - combined samples

The previous sample sizes were rather small. We also performed the Mann Whitney U test on the combined samples, using the verb and preposition combination for the target instances. The data collected was still specific to the verb and preposition, but the Mann Whitney U test was conducted on the scores of the combined samples. The result was, on this occasion, statistically significant. The group size and sum of the ranks are shown in table 5.6 in the columns headed *on & at*. The probability of getting the sum of the ranks for the positive group at 86 or above was 0.03. This was a significant result. However, it was the only one of the LO experiments that was.

5.8.5 The MDL Method: using ATCMs

The MDL method was applied to identification of the causative alternation using the data in lexicon A. Initially, this was performed with ATCMs. This followed earlier work reported in (McCarthy & Korhonen, 1998). In the work described in this paper, a sample of 30 verbs was used. This was the same sample used for the WSD experiment on page 65 of chapter 3. These verbs were selected at random, and did not necessarily have the required SCFs. Half of the 30 verbs used in the experiment reported in this paper were removed from consideration by virtue of not having the required SCFs. Those removed were identified by the human judges as non-participating verbs. Using the MDL method of diathesis detection on the 15 remaining verbs provided an accuracy of 87%. This was very encouraging when compared to the random baseline of 50%.[14] Although these results were encouraging, we show here that there are problems with using description lengths, and in particular with using ATCMs.

For the ATCMs, a prior model for $p(c)$ is required in the calculations. When obtaining models for individual slots, the prior has usually been collected from the target slot (Resnik, 1993a; Abe & Li, 1996; McCarthy, 1997), although Ribas (1995a) experimented with different priors. When obtaining the description length for the combined model the choice of prior is not obvious since the conditional data, dependent on the verb, comes from two slots. For alternating verbs, the data from the two slots should be similar. The prior however is not specific to the verb, and combining the data from both slots may confuse the results. A solution might be to take a prior from the slot that is considered the 'base form' in a directed alternation rule. However, the issues of directionality

[14] The random baseline reflected the two way decision for participation.

Table 5.7: The effect of the prior on ATCM results for the MDL method

prior	accuracy
object	11/22 (50%)
subject	17/22 (77%)
obj and subj	12/22 (55%)
all nouns	10/22 (45%)

are far from clear cut. In McCarthy & Korhonen (1998), the prior was obtained from data at the subject slot. However, there is as much justification for using the object slot.

We present here the results that we obtained when we repeated the experiment using a slightly larger set of 22 verbs. This set included the 15 verbs used in McCarthy & Korhonen (1998). The sample was extended to include more positive examples, since in McCarthy & Korhonen (1998) the majority of candidates (10 out of 15) were negative. The 22 verbs we used were:

positive:
begin, *break*, *change*, *drop*, *end*, *grow*, *move*, *ring*, *swing*, *worry*.

negative:
add, *ask*, *believe*, *charge*, *choose*, *cut*, *eat*, *expect*, *feel*, *help*, *know*, *like*.

We used four different priors. These were obtained using data from:

1. the object slot
2. the subject slot (as in (McCarthy & Korhonen, 1998))
3. both the subject and object slot
4. all nouns in our corpus sample

The results are shown in table 5.7.

One false positive arose because *cut* was identified in every case as taking the causative. This is easily explained since *cut* takes the middle alternation. The middle alternation is exemplified by 30, and is a close relative of the causative. The adverbials which are characteristic of this alternation are dropped by the SCF acquisition system. Thus the predicate was misclassified.

(30) a. the butcher cuts the meat.

b. the meat cuts easily.

The experiments with priors using the object slot, combined object and subject slots and the all nouns sample, all suffered from false positives. When we used the prior from the subject slot, the errors were all false negatives, with the exception of *cut* which is explained above. Ribas (Ribas, 1995a) and Li[15] have both observed that the association score is greatly affected by changes to the

[15]Personal communication.

prior. In this experiment, the prior from the subject slot diminished the effect of the overwhelming frequency of the **person** class. The subject slot prior had a high probability at the **person** class and this reduced the association score for this class. Since the association score for this highly populated class was reduced, less spurious similarities were found across the target slots.

The results for the ATCM were so dependant on the prior used that we turned our attention to other TCM types. LLRTCMs did not present as a good choice because the data description length does not relate to the number of bits required for encoding the data. In these models, LLR is used as a heuristic, in place of the data description length, From a small experiment using the 15 verbs used in McCarthy & Korhonen (1998), all candidates were rejected from participation, apart from one false positive.

We turned our attention instead to the PTCMs. The description lengths of these are more clearly related to the number of bits for description, and so to MDL.

5.8.6 The MDL Method: using PTCMs

Using a larger sample of verbs, we investigated how well the causative alternation could be identified using the probabilistic models and the MDL method as before. We also extended the experiments to include the conative alternation.

A problem for diathesis alternation detection using semantic preferences is that many slots in different frames have similar preferences even where they do not alternate. We referred to this problem, and gave an example, on page 130 above. There will be some differences in the preferences at the corresponding slots of non-alternating verbs, but these may be small. One possibility for increasing precision on the task, whilst reducing coverage, is to filter out verbs where the alternating slots co-occur in the same SCF and the slot fillers are similar. Our example of this is the transitive frame of the causative alternation. We experimented with and without filtering out the verbs which showed similar preferences at the subject and object slots of the transitive frame.

Causative:

We used the sample of 110 verbs (54 positives and 56 negatives) which were used in the lemma-based experiment on page 141. These sets are listed again here to contrast them with the set of verbs filtered because the subject and object slots of the transitive SCF had similar preferences, using PTCMs and the MDL method.

Positive sample (54 verbs):

bake, *begin*, *bend*, *blend*, *boil*, *break*, *burn*, *calm*, *change*, *clear*, *close*, *continue*, *cook*, *crack*, *crash*, *decrease*, *develop*, *drive*, *drop*, *dry*, *end*, *expand*, *finish*, *flood*, *fly*, *grow*, *hang*, *hurt*, *improve*, *increase*, *land*, *march*, *match*, *melt*, *mix*, *move*, *open*, *record*, *repeat*, *ring*, *roll*, *settle*, *shut*, *sink*, *split*, *spread*, *start*, *stop*, *stretch*, *swing*, *train*, *turn*, *vary*, *wake*.

Negative sample (56 verbs):

add, *admit*, *answer*, *ask*, *attack*, *believe*, *bother*, *catch*, *charge*, *choose*, *climb*, *compare*, *cost*, *cut*, *declare*, *demand*, *dress*, *drink*, *eat*, *expect*, *feed*, *feel*, *hear*, *help*, *hide*, *imagine*, *imply*, *investigate*, *kick*, *know*, *like*, *live*, *love*, *miss*, *nod*, *observe*, *pack*, *pass*, *pay*, *perform*, *plan*, *pull*, *read*, *remain*, *remember*, *shout*, *sing*, *steal*, *survive*, *suspect*, *think*, *understand*, *wash*, *win*, *work*, *write*.

Verbs from either of the above groups filtered out because the subject and object of the transitive are similar (71 verbs):
admit, *answer*, *ask*, *attack*, *bake*, *believe*, *bend*, *blend*, *boil*, *bother*, *burn*, *calm*, *catch*, *charge*, *clear*, *climb*, *cook*, *crack*, *crash*, *cut*, *declare*, *decrease*, *dress*, *drink*, *drive*, *drop*, *dry*, *eat*, *expand*, *expect*, *feed*, *finish*, *flood*, *fly*, *grow*, *hang*, *hear*, *help*, *hurt*, *imply*, *kick*, *know*, *land*, *like*, *love*, *match*, *melt*, *miss*, *mix*, *move*, *pack*, *pass*, *pay*, *pull*, *remain*, *remember*, *ring*, *roll*, *sink*, *split*, *spread*, *stop*, *stretch*, *suspect*, *swing*, *think*, *train*, *turn*, *understand*, *vary*, *wake*.

The filtering removed a surprisingly high proportion of verbs. Cases where similar sets of lexical items appeared in both grammatical slots in the transitive frame were filtered as anticipated. In other cases, verbs were removed where there were in fact differences between the selectional preferences at the two slots. Frequently this was because the actual semantic preferences of the verbs for the different slots were concerned with the same particular area of WordNet. Since the whole TCM was considered, differences in one particular area were sometimes drowned out by similarities in many other areas. For example, *eat* showed selectional preferences for **person** at the subject slot, and **food** at the object slot. These classes are both beneath the **entity** class which is only one of the eleven roots of WordNet. When the argument head data was combined, there was a high probability at the **entity** class vicinity and the system incorrectly identified *eat* as taking the causative alternation. Moreover, in some cases our acquired preferences were not discriminatory enough. The TCMs were not always low enough for a distinction to be made between important subclasses. For example, for *melt, cook, burn* and *boil*, the distinction between **person** subject slot and **substance** object slot was lost by a tree cut above these classes.

The remaining 39 verbs after filtering were:
add, *begin*, *break*, *change*, *choose*, *close*, *compare*, *continue*, *cost*, *demand*, *develop*, *end*, *feel*, *hide*, *imagine*, *improve*, *increase*, *investigate*, *live*, *march*, *nod*, *observe*, *open*, *perform*, *plan*, *read*, *record*, *repeat*, *settle*, *shout*, *shut*, *sing*, *start*, *steal*, *survive*, *wash*, *win*, *work*, *write*.

Of these there were 16 positive candidates and 23 negative candidates. The errors are shown below:-

1. false positives: *develop*, *record*, *steal*, *wash*
2. false negatives: *continue*, *end*, *march*, *settle*, *start*

The filtering process resulted in a marked increase in accuracy, although of course at the expense of coverage (see table 5.8). We abandoned it for the rest of our experiments because of the reduction in coverage.

Conative:
This alternation is shown in example (25) above. The test sample was the same as that which we used for the LO experiments in section 5.8.4 above. As before, we evaluated performance in two experiments which were specific to the prepositions *at* and *on*.

The results are shown in table 5.9. The table gives the breakdown between true positives (TPs), true negatives (TNs), false positives (FPs) and false negatives (FNs). The accuracy for the conative

Table 5.8: Filtering out difficult candidates

	accuracy	sample coverage
Without Filtering	63%	100%
With filtering	77%	35%

Table 5.9: Conative results

	on	*at*
TPs	bang cut press pull	pull push shoot tug
TNs	work	
FPs	agree move remain	call move remain work
FNs		

on sample was 62%. All errors were false positives and moreover only one negative case was identified. The abundance of false positives arose because the PP frame, specific to *on*, had such a low frequency compared with the transitive frame. The direct object data swamped the PP data when the data at the target slots were combined for the MDL method.

In addition to the problem of disparate relative frequencies of the target slots, some verbs, such as *agree*, take the same noun in the PP and transitive frames, see for example (31) below, without this being a case of the conative alternation. Levin (1993) observed that the conative alternation has specific semantic constraints. The verb in the intransitive PP frame describes an attempted action without specifying whether the action was actually carried out. Participating verbs involve notions of contact and motion. Two new informants, not our original judges, classified *agree on* positively as taking the alternation. Making distinctions between verbs like *push*, which does participate in the conative, and verbs such as *agree*, which does not, requires knowledge of the subtle semantic prerequisites for the alternation. The new informants were not aware of the semantic prerequisites specified by Levin. The original two informants were. The automatic method likewise is not sensitive to these prerequisites. It simply identifies cases where the slot fillers seem to be capable of switching position. It may be that some additional a priori knowledge would improve accuracy for identifying alternations with specific semantic constraints. Specification of the semantic properties of the verbs would help only if the semantic properties could be detected automatically. The semantic properties of the arguments could be detected automatically, using acquired selectional preferences such as ours. However, the semantic properties of the arguments vary depending on the participating verb, and do not usually form a coherent semantic type across all participating verbs.

(31) a. They agreed the cost.

b. They agreed on the cost.

Accuracy for conative *at* was only 50% (the same as our random baseline). No negative decisions were made by the system. When the outcomes of these two experiments were put together,

Alternation	Average Frequency Ratio
Causative	1.16
Conative 'on'	28.99
Conative 'at'	32.72

Table 5.10: Average frequency ratios

the accuracy obtained for the full 16 candidates was $\frac{9}{16} = 56.25\%$. This was not a very encouraging result.

Relative Frequencies and the MDL Approach

Comparing the description length costs of the TCMs relied on the alternating slots having similar frequencies. Lapata (1999) referred to verbs which have similar frequencies of alternating SCFs as having a high degree of typicality. The typicality of a verb for an alternation will depend on the verb and the alternation. We calculated an average frequency ratio for each of the alternations we experimented with. This used the frequency ratio between the frames for a given alternation, averaged over all verbs, as shown in equation 5.24. The calculation is specific to the alternation (X). We took the average for all candidate verbs in the test sample (verbs) of the ratio between the most frequent frame (SCF1) for the alternation and the less frequent frame (SCF2). SCF1 and SCF2 were determined with respect to the alternation; they were not verb specific. The average frequency ratio is minimised when the alternating SCFs have equal frequencies. At the limit, this gives a value of 1. The value increases as the difference between the frequencies of the alternating SCFs increases.

$$\text{average frequency ratio}_X = \frac{\sum_{v \in verbs} \frac{\text{freq}(v, \text{SCF1}_X)}{\text{freq}(v, \text{SCF2}_X)}}{|verbs|} \tag{5.24}$$

The average frequency ratio for the causative and conative alternations are given in table 5.10.

The ratio for alternations which involve specific prepositions, such as the conative, was high (not close to 1). This explains the poor performance using the MDL technique for these alternations. The MDL technique is expected to work for verbs where the alternation is typical. Nevertheless, there are many verb and alternation combinations which have disparate frequencies between the alternating frames, yet are provided as exemplars for the alternation by Levin. For example, *cut at* is provided as a prime example of the conative by Levin (1993, p41), however the ratio between the transitive frame and PP_at frame was $\frac{657}{4}$ in lexicon A. This suggests that Lapata's choice of the 'typicality' terminology was not altogether appropriate

We now turn to the similarity approaches, which permit diathesis detection regardless of the typicality of an alternation.

5.8.7 The Similarity Approach - Comparing Probability Distributions

In this subsection, we compared the selectional preferences at the target slots, represented as probability distributions, for diathesis alternation identification. The results discussed here are obtained from lexicon D. It was hoped that a wider variety of alternations could be covered with a larger

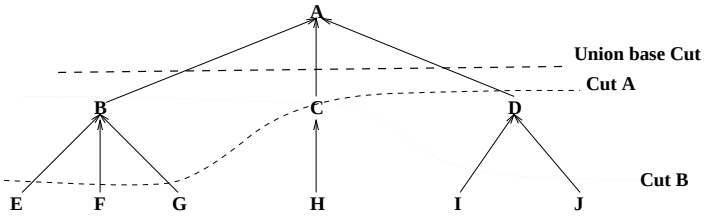

Figure 5.3: A union base cut

quantity of data and a more accurate parser. However, out of the alternations in table 5.4, only the locative preposition drop (Levin 1.4.1) had sufficient candidates meeting the frequency threshold to be included for experimentation. This alternation behaves as in example 32.

(32) a. She crossed over the river.

b. She crossed the river.

The similarity approach compares the PTCMs at the target slots of the SCFs specified by the Levin–SCF mapping. The measures of distributional similarity which we discussed in section 5.5.3 require discrete probability distributions over the same set of items. PTCMs cover the leaves of WordNet, but can do so by cutting across different levels of the WordNet hierarchy. Before application of a similarity measure, the two probability distributions must be mapped to a common set of classes for comparison. To do this we used a base set of classes across WordNet. Two probability distributions over this base cut were then produced from the original PTCMs at the target slots. This was performed by using the method outlined on page 33 in chapter 2 for finding the probability of a class from the estimates on a PTCM above or below this class.

We used two different ways of identifying the base classes. The first was simply to take a base cut at the eleven root classes of WordNet. We refer to this as the 'root base cut'. The other method that we used was to produce a base cut from classes of the two PTCMs. This was obtained by taking all classes from the union of the two PTCMs which were not subsumed by another class in this union. Duplicates were removed. This is termed the 'union base cut'. A union base cut is illustrated in figure 5.3 for an imaginary hierarchy. This union base cut is obtained from the union of two cut models, A and B, given in this diagram without probabilities. A new PTCM is then produced from each original PTCM. The new PTCM contains the classes on the union base cut. The probabilities for these classes are calculated using those on the original PTCM. The probability for a superordinate class is obtained by combining the probability estimates for all of its hyponyms on the lower cut. The probabilities for the original PTCMs A and B are shown in figure 5.4 alongside the new PTCMs formed with the union base cut.

In these experiments, we based the gold standard for evaluation on the decisions of four human judges. All judges were given a list for each alternation for which we had sufficient candidates from the data. These alternations were:

- causative
- conative (*at* and *on*)

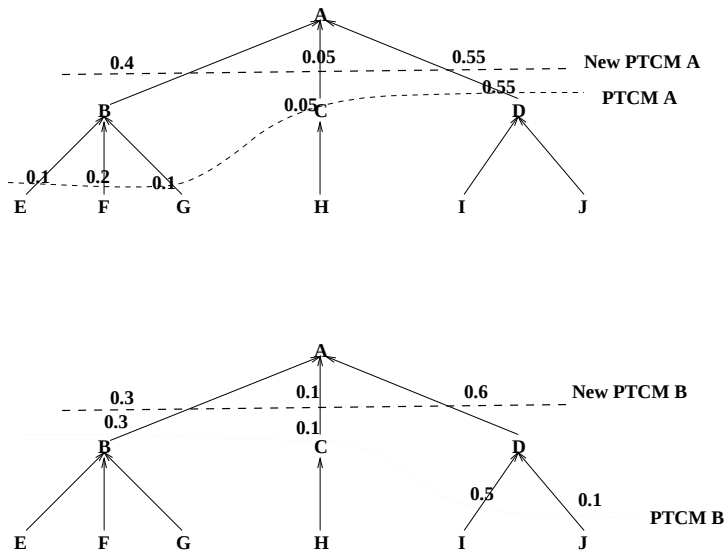

Figure 5.4: New PTCMs at the union base cut

Alternation	kappa	z	significance (p)
Causative	0.71	12.4	0.00003
Conative on	0.60	3.1	0.001
Conative at	0.67	3.5	0.00023
Locative preposition drop	0.35	1.6	0.0548

Table 5.11: Human agreement

- locative preposition drop

The verbs were sorted alphabetically in these lists. The judges had to decide on participation for each candidate. The 'don't know' category was permitted. The kappa statistic was calculated for each alternation and the results are shown in table 5.11. The judges showed significant levels of agreement on all alternations except for the locative preposition drop. For our experiments we required at least 3 candidates in each category (positive or negative) for the Mann Whitney U test. The locative preposition drop had 3 candidates in each category with more than 75% agreement. This alternation was not used because the sample was so small and because the level of agreement between judges was not significant.

The mean was used as a threshold on similarity scores for determining accuracy. The median was also used. When calculating the mean and accuracy, we ensured there were an even number of positive and negative candidates after verbs without sufficient agreement were removed. This was done by randomly removing the surplus number of candidates from the larger category.

Causative

118 candidates were selected from the data. These were evenly split between positive and negative candidates, according to us. After removing those with less than 75% agreement we had 46 positives and 53 negatives remaining. Seven of the negative candidates were selected at random and removed from the negative sample before this was used for determining accuracy using the mean as a threshold. The candidates remaining were:

positive sample (46 verbs) :

accelerate, *bang*, *bend*, *boil*, *break*, *burn*, *change*, *close*, *cook*, *cool*, *crack*, *decrease*, *drop*, *dry*, *end*, *expand*, *flood*, *fly*, *improve*, *increase*, *land*, *march*, *match*, *melt*, *open*, *repeat*, *ring*, *rip*, *rock*, *roll*, *shatter*, *shut*, *slam*, *smash*, *snap*, *spill*, *split*, *spread*, *start*, *stop*, *stretch*, *swing*, *terminate*, *tilt*, *turn*, *wake*.

negative: (46 verbs)

add, *admit*, *answer*, *ask*, *attack*, *believe*, *borrow*, *catch*, *choose*, *climb*, *cost*, *declare*, *demand*, *drink*, *eat*, *expect*, *feel*, *help*, *imagine*, *kick*, *knit*, *know*, *miss*, *notice*, *outline*, *pack*, *paint*, *pay*, *perform*, *plan*, *practise*, *prescribe*, *proclaim*, *pull*, *read*, *remain*, *remember*, *sing*, *steal*, *suck*, *survive*, *understand*, *warn*, *wash*, *win*, *write*.

Table 5.12 shows the results for detection of the causative alternation using the four similarity measures described in section 5.5.3. The root base cut was used in all cases and the PTCMs at the target slots were produced without WSD on the input data. The sample was large enough for a z score to be obtained from the ranks of the Mann Whitney U test. The z score was looked up in a table of the normal distribution to provide the probability (p) of obtaining the score by chance, i.e. if there were no relationship between the similarity measure and participation. Values of p less than the 0.05 [16] significance level indicated that there was a significant relationship between the similarity measure and participation. Values of p less than 0.01 were highly significant. The results for all similarity scores for the causative were all highly significant. The final two columns in the

[16] This represents a 95% confidence level for a one-tailed significance test.

	Mann Whitney z	significance (p)	mean	median
ED	-4.2	0.0003	72	65
cosine	3.7	0.00011	65	63
L1 norm	-4.05	0.0003	68	63
αSD	-4.03	0.0003	71	63

Table 5.12: Causative identification with 4 similarity measures

table show the accuracy obtained when we used the mean and median respectively as thresholds to determine participation. Accuracy exceeded the 50% baseline in all cases. Performance was best for ED, with αSD a close second. These two measures were used in further experiments with the causative. The mean outperformed the median in all cases.

Causative results with more specific cut models

In section 5.8.6 on page 146, we observed that selectional preferences at different grammatical slots often involved the same area of WordNet. This typically occurred below the WordNet root **entity** which encompasses classes such as **person**, **object** and **food**. The root base cut could not differentiate the probability distributions in these cases because only the probability estimates at **entity** were compared.

We investigated the effect of (i) using a more specific base cut and (ii) WSD of the input data. We experimented using the union base cut and the root base cut without WSD, and also using each of the three WSD options described in chapter 3. The results for the ED similarity measure are displayed in table 5.13. Those for the αSD similarity measure are given in table 5.14. The mean performed better than the median in most cases, but not all. For ED the results were rather disappointing. WSD and the union base cut made matters worse rather than better. When αSD was used, the combination of the union base cut and FirstS WSD did improve matters a little. However, these differences in accuracy were not significant, using the chi-squared test.[17]

Conative

The conative sample was rather small after disregarding those verbs with less than 75% agreement between judges. If we had separated the conative experiment into two preposition specific experiments we would have had just the statutory number for the small sample size of the Mann Whitney U test (we only had 3 positive candidates). Instead we combined the data using positive and negative candidates with specified prepositions. The four similarity scores were used with the root base cut and no WSD as before. The results are given in table 5.15.

There was a significant relationship between all of the similarity scores and participation. However, the difference between accuracy and the baseline was not significant (on the chi-squared). This was due to the small sample size.

Conative results with more specific cut models

For completeness we show the results obtained using the union base cut and the WSD options as before. The results using ED are in table 5.16 and those for αSD are in table 5.17. WSD increased

[17]The chi-squared test was used because we were comparing frequencies.

root base cut				
	Mann Whitney z	significance (p)	mean	median
NOWSD	-4.2	0.00003	72	65
SPass	-2.75	0.003	62	63
FirstS	-3.29	0.0005	67	63
COMB	-3.48	0.00023	68	65
union base cut				
	Mann Whitney z	significance	mean	median
NOWSD	-4.59	0.00003	72	70
SPass	-3.10	0.001	64	63
FirstS	-3.11	0.0009	65	67
COMB	-2.81	0.0025	65	65

Table 5.13: Identifying the causative using ED with WSD options

root base cut				
	Mann Whitney z	significance(p)	mean	median
NOWSD	-4.03	0.0003	71	63
SPass	-3.08	0.001	66	63
FirstS	-3.5	0.00023	70	61
COMB	-3.7	0.0001	67	67
union base cut				
NOWSD	-4.3	0.00003	73	70
SPass	-1.9	0.0287	61	57
FirstS	-4.4	0.00003	75	67
COMB	-3.2	0.0007	64	61

Table 5.14: Identifying the causative using αSD with WSD options

	Mann Whitney sum	significance (p)	mean	median
ED	30	0.09	83	67
L1 norm	29	0.07	83	67
cosine	54	0.008	67	83
αSD	26	0.02	67	83

Table 5.15: Conative identification with 4 similarity measures

root base cut				
	Mann Whitney sum	significance	mean	median
NOWSD	30	0.09	83	67
SPass	29	0.07	67	67
FirstS	23	0.004	75	83
COMB	26	0.02	75	83
union base cut				
NOWSD	33	0.197	67	67
SPass	32	0.1548	67	67
FirstS	26	0.0206	75	83
COMB	27	0.0325	75	67

Table 5.16: Identifying the conative using ED with WSD options

root base cut				
	Mann Whitney sum	significance	mean	median
NOWSD	26	0.02	67	83
SPass	26	0.02	75	83
FirstS	26	0.02	67	83
COMB	22	0.002	83	83
union base cut				
	Mann Whitney sum	significance	mean	median
NOWSD	34	0.2	58	67
SPass	26	0.02	67	83
FirstS	34	0.2	58	67
COMB	22	0.0022	83	83

Table 5.17: Identifying the conative using αSD with WSD options

accuracy when αSD was used, however these differences in accuracy were not significant, using the chi-squared test.[18]

Error Analysis for the Mean and Median Thresholds.

The results obtained were dependent on the thresholds taken. On the whole, performance was better for the mean than the median for the causative experiments, but there were cases where the converse was true. For the conative experiments, the median outperformed the mean in many cases.

It is interesting to look at the effect of the threshold used on the types of errors made. Errors were classified as false positives or false negatives. False positives arose when a non-participating verb was wrongly identified by the system as taking the alternation. False negatives occurred when a participating verb was not identified as taking the alternation by the system. When the

[18]The chi-squared test was used because we were comparing frequencies.

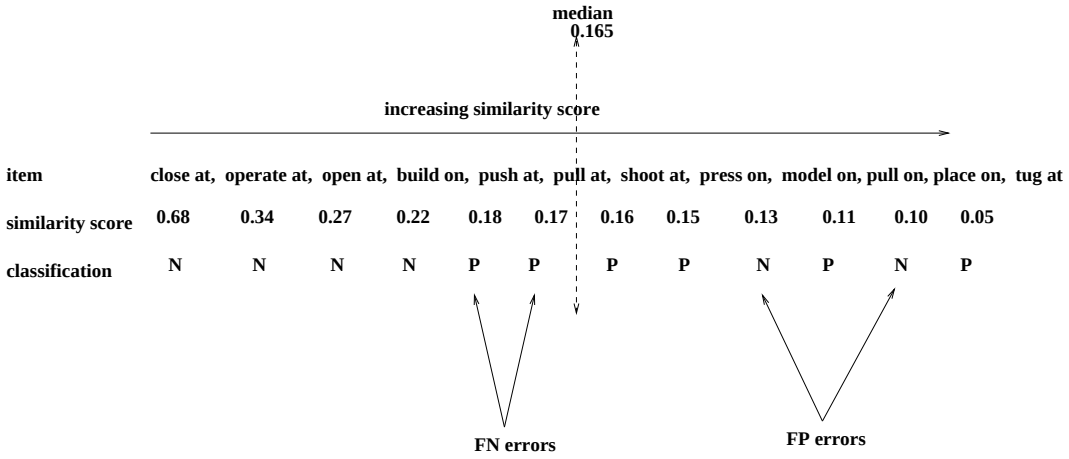

Figure 5.5: Using the median as a decision point

median was used as a threshold, the number of false positives and false negatives were evenly balanced. This is because the median threshold is, by definition, taken midway between the test items arranged in order of their similarity scores. There were an even number of items on either side of the decision point, and an even number of positive and negative candidates in our test sample. Thus, the errors on either side of the decision point were equal in number. This scenario is illustrated in figure 5.5, the data is taken from the conative experiment using ED with the root base cut and no WSD. The decision of the judges is indicated in the row marked classification, P for positive and N for negative.

Across the entire set of experiments, there was typically a larger number of false positives than false negatives when the mean was used as the threshold. The breakdown between error types is displayed in table 5.18 for a subset of our experiments. This is shown for experiments detecting the causative and conative alternation with the ED similarity measure, using the root base cut and no WSD, but was typical of all the experiments. The table gives the threshold, number of false positives, number of false negatives and accuracy when using both the mean and median. The mean usually produced a higher accuracy than the median, but gave rise to an increase in false positives. The mean was typically higher than the median for all measures except the cosine. The cosine is a true measure of similarity, as opposed to a measure indicating dissimilarity, and for this score the mean was lower than the median. These results indicate that the scores were not normally distributed, since in a normal distribution the mean and median are the same.

The polysemy of the verbs may be one explanation for the large number of false positives. The SCFs and data of different senses should ideally not be combined, at least not for coarse grained sense distinctions. We tested the false positive and true negative candidates to see if there was a relationship between the polysemy of a verb and its misclassification. The number of senses (according to WordNet) was used to indicate the polysemy of a verb. The Mann Whitney U test was performed on the verbs found to be true negative and false positive using the root base cut and no WSD options. A significant relationship was not found between polysemy and misclassification

alternation	threshold type	threshold	accuracy %	num FPs	num FNs
causative	median	0.23	65	16	16
causative	mean	0.28	72	19	7
conative	median	0.16	67	2	2
conative	mean	0.21	83	2	0

Table 5.18: Error analysis on experiments using ED, no WSD and the root base cut

5.9 Summary and Conclusions

Diathesis alternations are systematic variations in the syntactic realizations of verbs. They are systematic in that the same alternation occurs for a number of verbs having some similar semantic component. In RSAs, a particular semantic role will occur in different grammatical slots in the alternating realizations. These alternations are neatly exemplified using the same lexemes in alternating variants, as in example (33).

(33) a. She began the meeting.

b. The meeting began.

Detecting genuine participation is not straightforward. Automatic parsing technologies make it possible to identify instances of the alternating frames. For some alternations (the dative and benefactive) this is sufficient for correct identification. However, due to sparse data the same lexemes may not occur in both variants in the corpus data at hand. Moreover, some lexical items cannot occur in both variants, even for verbs which do participate. The anticipated difficulties are corroborated by evidence from the lemma-based method described in section 5.5.4 with experimental results presented in section 5.8.4.

In this chapter, we described and demonstrated two methods for observing participation in RSAs using selectional preferences. The preferences provided generalisations over the lexical fillers given any particular SCF and slot combination.

The two methods were the MDL method and the similarity approaches. The MDL method relies on a comparison between the cost of separate models for the alternating slots and the cost of a model if the data is combined. This showed encouraging results for the PTCMs, but did not work well for verb and alternation combinations with substantial differences in the frequencies of the alternating frames.

The similarity measures are more generally applicable. We discovered a highly significant relationship between measures of distributional similarity and participation in the causative and conative alternations. We obtained a rudimentary idea of the level of accuracy by using thresholds to determine participation. The arithmetic mean and median thresholds were used. If such thresholds were to be applied in earnest then they would need to be obtained from held out data. The mean produced slightly better results for the causative, but a higher proportion of false positives. Accuracy of 72% was obtained for the causative alternation using the ED score. This was significantly above the baseline of 50%. Accuracy for the conative alternation was 83% but the difference compared to the baseline was not statistically significant owing to the small sample size.

Neither the greater specificity arising from WSD of the input data nor comparing cuts at a more specific level (the union base cut) improved matters significantly.

The most important impediment to using these methods for discovering participation is sparse data. Larger corpora are required if we are to use this method with a wider set of alternations, and with less common verbs. Our use of automatic methods for producing the SCF lexicon and preferences allows this with human effort only required for selecting and tagging candidate verbs for evaluation.

There are other possibilities for overcoming the sparse data problem, other than the corpus size. Briscoe & Carroll (1997) identified the statistical filter as the main source of error in their SCF acquisition system. More attention to this component could substantially increase the quantity and quality of the candidates selected.

An important characteristic of all the alternations in table 5.4 on page 138 is that PPs are involved in at least one alternating SCF. The preferences for these slots are acquired with reference to the specific preposition. Thus, in the conative experiments we have considered candidates with respect to a particular prepositions (*on* or *at*). This drastically reduces the quantity of data available for PP frames for the candidate verbs. There may be ways around this. It may be worth investigating if grouping prepositions would help for alternations involving PPs. [19] The prepositions could be grouped by hand, or clustered automatically using distributional evidence.

False positives were the largest source of error in our experiments. Although we did not find a significant relationship between polysemy and misclassification, it may be that our experiment was not successful in isolating the types of verb sense that give rise to the false positives.

Aside from looking at verb senses, another strategy which might reduce the false positives is to make use of more stringent criteria for diathesis alternation detection. Semantic criteria may be useful, however, semantic properties of the verbs would be difficult to detect automatically. Semantic properties of the arguments vary depending on the participating verb, and do not usually form a coherent semantic type across all participating verbs.

We believe it will help to look at sets of alternations collectively, rather than one alternation at a time. Levin (1993) identified classes of verbs which participate in particular sets of alternations. She also indicated constructions which correlate (positively or negatively) with other diathesis patterns. For example, she pointed out that the 'X's Way Construction' (Levin, 7.4) does not occur for unaccusative verbs. If the evidence for a particular verb is combined, using observed alternations, then we might predict unobserved alternations using Levin's classification. We would of course need more observed evidence. The alternations in table 5.2 on page 135 require more syntactic evidence for identification. Evidence at the phrase level is not currently retained in our SCF lexicon. However, the SCF acquisition system could be modified so as to include it. Syntactic evidence, such as possessive markers, should be reliably identified. These could narrow the search for participants for these alternations considerably. Perhaps to the extent already seen with the benefactive and dative alternations, where semantic evidence is not even required.

Our method of detecting participation could be used alongside other, complementary cues. For example, Resnik (1993a) showed the relationship between the strength of the selectional preference of a verb and its participation in object drop alternations. These alternations are the un-

[19] One example where this might work is *put* which subcategorizes for a locative PP. It remains to be seen whether enough other verbs also subcategorize for PPs headed by a preposition class.

expressed object alternations (Levin 1.2). One could use this relationship alongside our methods, which identify RSAs. Together, methods such as these could be used for identifying the correct classification of a verb within Levin's taxonomy.

Chapter 6

Conclusion

This chapter is organised into two sections. The first section summarises the contributions of this thesis and the second section outlines directions for future research.

6.1 The Contributions of this Thesis

The main contribution of this thesis has been to show how automatically acquired SCFs and selectional preferences can be combined to predict verbal participation in diathesis alternations (chapter 5). This is an important contribution to NLP research because diathesis alternations lie at the bridge between syntax and lexical semantics. Information about participation can be used to predict unseen subcategorization behaviour and to help classify verbs semantically (Levin, 1993).

To identify potential candidates from their syntactic behaviour, we used the SCF acquisition system of Briscoe & Carroll (1997). This provides verbal entries classified according to an inventory of 163 SCF classes. Each entry includes the argument head data found at each slot in the training data for the given verb and SCF combination. For selectional preference acquisition, we modified a system for acquiring preferences as tree cut models (TCMs) across WordNet, originally devised by Li & Abe (1995, 1996). We experimented with identification of proper nouns (as described in chapter 2) and automatic WSD of the argument head data (as described in chapter 3). Previous systems have, on the whole, acquired preferences with hand disambiguated data (Ribas, 1995a), disambiguation assisted by an MRD (Pozanski & Sanfilippo, 1996) or, more commonly, no WSD (Resnik, 1993a; Li & Abe, 1998; Abe & Li, 1996; Grishman & Sterling, 1993; Pereira et al., 1993; Rooth et al., 1999). Our system can be run without any WSD, or with two distinctive WSD options. The first uses a first sense heuristic (FirstS) and the second uses an iterative approach (SPass), where the preferences acquired from the raw data are used for subsequent WSD on a second iteration. Others have also investigated iterative approaches (Abney & Light, 1999; Clark & Weir, 1999). Our system can be run using a combination of FirstS and SPass (COMB).

We reported several formal evaluations of the selectional preference acquisition system in chapter 4. The evaluations were performed to compare our preference models to those produced by other researchers, and to compare the affect of various parameter options. We performed the following formal evaluations:-

1. Evaluation against the selectional restrictions provided in LDOCE. This evaluation was problematic in three respects. Firstly, LDOCE provides hard constraints on the selectional properties of verbal entries, whilst our system used the frequency information available from the corpus data to provide a system of selectional preference on a continuum. To overcome this, we applied a threshold on our preference scores and compared the result to the LDOCE constraints, although much information is, of course, lost in this way. Secondly, automatic preferences cannot possibly cover phenomena that were not attested in the training corpus: this happens frequently since LDOCE lists rare and specialised senses. Thirdly, the automatic preferences were penalised where they related to genuine preferences which were omitted by the lexicographers. Despite these obvious shortcomings, we performed the evaluation to find out the overlap between the automatically acquired preferences and ones specified a priori by lexicographers. Mismatches of the second type, where LDOCE restrictions were not found in the corpus data, affected recall. Mismatches of the third type, where TCM preferences were not recorded in LDOCE, affected precision. The precision and recall figures showed the extent to which our parameter settings increased the range of preferences observed or were more conservative and covered only the more salient preferences.

2. Evaluation against examples provided in CIDE. We calculated the proportion of the dictionary examples, listed for a sample of verbal entries, which were covered by the TCMs. For this evaluation, the dictionary examples were covered by the TCMs if the argument head at the specified slot in the example belonged (directly or indirectly via a hyponym relationship) to one of the classes on the TCM above a stipulated threshold. This was compared to a baseline which was determined by the average proportion of classes which have preferences above the threshold on the TCMs for the verbs in the dictionary examples. This baseline gave an indication of how discriminatory the TCMs were. This evaluation, like the LDOCE one, was problematic in that rare or specialised senses reported in the dictionary may not be attested with sufficient frequency in the training data, if at all. Nevertheless, it provided a way of comparing the behaviour of the different parameter settings.

3. WSD. This task was used because of the availability of appropriate test data (SemCor), and because we already had the machinery in place for WSD, since we used preferences in one of our options for WSD of the argument head data. We investigated how preferences performed at WSD, and how the various parameter options affected performance. A further reason for this method of evaluation, is that many other researchers have also applied automatically acquired preferences to the WSD task. Direct comparison of reported results is fraught with difficulties because different test and training data have often been used. However, in addition to the SemCor evaluation, we entered our preference acquisition system, with one parameter setting, for the SENSEVAL competition (Kilgarriff et al., 1998). Our system (Carroll & McCarthy, 2000) had a similar performance to the only other system (Kilgarriff & Rosenzweig, 2000) that used selectional preferences alone.

4. Pseudo-disambiguation. Many of the proximity-based semantic classifications, which can be used as selectional preference models, have been evaluated on the task of choosing between genuine and artificially produced word pairs. Our system performed less well than many systems which have used automatically produced classifications (Pereira et al., 1993; Rooth et al., 1999), apart from Grishman & Sterling (1993). Grishman & Sterling obtained a very low recall (34%) set against a low error rate (9%), however they evaluate using erroneous and correct parses. It

is therefore inappropriate to compare our results with those of Grishman & Sterling because of differences in the task. Direct comparison with the results of Pereira et al. and Rooth et al. is awkward because of substantial differences in the task. Our preferences were obtained using substantially smaller quantities of training data, with no frequency threshold applied to the lemmas involved and were tested on a different test set. It is likely that, even given the differences in the training and test data, preferences acquired using a manmade classification, like WordNet, will be less accurate than those acquired within an automatically constructed classification (Li & Abe, 1996).

This task did not provide any significant differences for any of the parameter options which were tried.

The LDOCE and CIDE evaluations have not been used in other research reported in the literature. We used them to highlight differences in the TCMs brought about by the various parameter settings, rather than to provide a figure of merit for the TCMs.

6.1.1 Modifications to the Selectional Preference Acquisition System

The following modifications were made to the basic approach devised by Abe & Li (1996):-

1. The creation of new leaves at internal classes. These were created for all hyperonym classes of WordNet. This was done so that all word senses fell under the classes on any TCM across WordNet. This avoided Li & Abe's strategy of pruning WordNet at classes where a word with direct membership of the class occurred in the argument head data. Their approach resulted in many TCMs being restricted to the WordNet roots at some point along the cut, because words at these roots, for example *entity* and *location*, occurred frequently in the BNC data.

Creating the new leaves gave a substantial reduction in the number of root cuts (TCMs at the dummy root which we created above the 11 WordNet roots). In a sample of 30 verbs, Li & Abe's strategy resulted in 6 root cuts (20%), whereas the strategy of using leaves for all internal classes did not give any root cuts for this sample. The effect of the latter strategy is to permit MDL to find the optimal level of generalisation, rather than restrict it to a rather shallow version of WordNet.

2. Differences in thresholding for the ATCMs. Alongside the previous modification, we changed the method of thresholding for the ATCMs so that classes from the prior model which were below the threshold were removed before calculation of the description length for the entire TCM. This was done to reduce the search space for efficiency purposes. It reduced the search space but, as a consequence, classes with a low prior probability were not considered for the ATCMs. This was not an ideal modification but one made to compensate for the additional search space encountered when avoiding Li & Abe's method of pruning. There was no significant difference in the precision and recall of the Li & Abe method on the WSD task, compared to our method. However, our method did allow a wider coverage. A class probability threshold was not applied when obtaining either the LLRTCMs or the PTCMs.

3. Log-likelihood ratio models (LLRTCMs). We added a new type of TCM which incorporated the binomial log-likelihood ratio test (LLR) for finding the optimal cut, and as a preference score on the cut models. This was a departure from MDL, since the description length no longer reflected the number of bits to describe the model and data. However, there is a relationship between MDL and the new models since the log-likelihood ratio can be used as a heuristic, in place of full MDL

methods (Dunning, 1993). We devised the LLRTCMs because LLR has been reported to be better at dealing with rare events than many other measures, including mutual information which the association score is based on.

The LLRTCMs had fewer root cuts than the PTCMs and ATCMs, giving better coverage of the data. For the verb and slot combinations where the ATCMs did provide preferences below the dummy root, the LLRTCMs were usually more conservative than the ATCMs. However, in cases where a verb has very strong preferences, the LLRTCMs were sometimes more specific and intuitive. The LLRTCMs were frequently more specific than the PTCMs. There were no significant differences in performance detected on formal evaluation between the ATCM, PTCM and LLRTCMs.

4. We experimented with the named entity recogniser of the GATE system (the VIE NE recognition system in version 1.1) for classifying proper nouns. The classified proper nouns were mapped to WordNet classes. ATCMs acquired with proper noun recognition were compared to ATCMs acquired using only the common nouns, and some pronouns. The proper nouns were discarded in the latter case. To our knowledge, this was the first use of software for recognising proper nouns within a selectional preference acquisition system.

Proper noun recognition increased the quantity of training data, which in turn increased coverage by reducing the number of root cuts. However, although on the SemCor evaluation, recall was improved slightly, precision (compared to the precision baseline) was reduced. These differences were not significant. For this reason, and because of the considerable computational cost of the proper noun recognition software and problems encountered when processing long sentences, the majority of work reported here did not use proper nouns. The accuracy of the named entity component, and robustness of GATE may well have improved in subsequent versions. For diathesis alternation detection, one of the main obstacles is data sparseness. Using proper noun data will help alleviate this and so further work in this direction is warranted.

5. For diathesis alternation detection, selectional preferences are required specific to the SCF, as well as to the slot and the verb. This is simple to achieve with the Briscoe & Carroll SCF acquisition system, since this provides the argument head data at the relevant slot within entries for a specified verb and SCF combination. For general evaluation of the selectional preference acquisition system, we evaluated TCMs specific only to the slot and verb. This made comparison easier with the other preference acquisition systems reported in the literature, since these do not stipulate the SCF. To see the effect of stipulating the SCF on preference acquisition, we compared selectional preferences acquired at the direct object slot of the `[np v np]` SCF with those acquired at the direct object slot generally, on the SemCor WSD task. There were no significant differences in recall or precision between the two sets of models observed on this task. The reduction in argument head data when stipulating the SCF was compensated for by the reduction in noise of the argument head data.

6.1.2 Selectional Preference Acquisition and WSD

WSD techniques were sought for disambiguation of argument head data which did not carry excessive computational demands for unsupervised training, nor any significant demands on human effort for supervised training. We required a disambiguation method applicable to as large a proportion of the argument head data as possible. It was felt that reliability and precision of the WSD

could be compromised in the interests of tagging a substantial proportion of the argument heads in a reasonable amount of time. This was acceptable because, for preference acquisition, WSD is performed over a set of argument heads collectively. We used two techniques in isolation:

1. FirstS– the first sense heuristic. This heuristic used the frequency estimates provided from SemCor. The first sense was chosen provided that three additional constraints were met. These limited the application of this heuristic to nouns with a clear predominant first sense and where the noun was not reported as being problematic for human taggers.

2. SPass– The selectional preferences acquired from the ambiguous argument head data were used to disambiguate the argument head data which was then input a second time to the selectional preference acquisition system.

These two techniques were also used together: the COMB option. For this option, the SPass technique was applied to nouns which did not meet the constraints for the FirstS technique.

The WSD options increased the homogeneity of the argument head data plotted in WordNet, typically this resulted in the MDL technique selecting more specific cuts. For verb and slot combinations where there were no strong preferences for the initial input data, the SPass technique could do little to improve matters. However, where there were clear areas of preference, these were reinforced by SPass. FirstS resulted in preferences being observed more readily. If there were any predominant collocates for a particular verb and slot combination, for example *open the door*, then there was a risk of the FirstS technique resulting in erroneous preferences. This happened if the FirstS technique chose the wrong sense for the noun in the collocation. It is safer to use the SPass technique for cases with strong collocations at a particular slot.

Generally speaking, there were no significant differences in precision and recall on the WSD and pseudo-disambiguation evaluations between the WSD options, including the option NOWSD which does not perform any WSD. However, coverage was increased by WSD because of the reduction in TCMs with cuts at the root. The FirstS option did improve precision and recall in situations with particularly sparse data, notably at the PP slot where we acquired preferences specific to the preposition.

Finally, we agree with Resnik (1997) that selectional preferences are not a panacea for WSD of nominal argument heads. However, they may help disambiguation when combined with other knowledge sources (Wilks & Stevenson, 1998b). Interestingly, verbal predicates are reported to be disambiguated quite well by their argument heads (Federici et al., 1999; Manning & Schütze, 1999; Stevenson, 1999). This would be worth considering further if disambiguation of verbs is to be investigated for diathesis alternation identification.

6.1.3 Diathesis Alternation Identification

In chapter 5, we demonstrated that automatically acquired SCFs along with selectional preference models can be used to establish whether a verb participates in a given alternation. Indeed, for alternations with distinctive SCFs, such as the dative and benefactive, the syntactic information alone was sufficient for identification. We demonstrated that, in other cases, such as the causative and conative, it is necessary to have selectional preference models as well. The TCMs provide evidence that arguments having a particular semantic type switch between different grammatical slots in the alternate realizations.

One method we proposed for incorporating the semantic information was to use the MDL costs, or description lengths, calculated when obtaining the TCMs. In this method, the description lengths for the TCMs at the alternating slots are combined and compared to the description length of the TCM that is obtained when the data at the separate slots is combined (the combined model). If the cost for the combined model is less than the combined costs of the separate models then participation is predicted. This worked for the causative alternation because the more homogeneous the data was, the cheaper the cost.

A significant problem with this method when using the ATCMs was that the results were radically affected by the choice of data for the prior model ($p(c)$). The LLRTCMs were also problematic for this method, since the LLRTCM cost is based on a heuristic and is not a true MDL description length calculated in terms of the number of bits needed to encode the model and data. There is no clear interpretation of a combination of costs from two separate LLRTCMs. Not surprisingly, applying this approach to the LLRTCM costs produced atrocious results.

The approach achieved reasonable results when applied to the PTCM costs for detecting the causative alternation. However, for other alternations, such as the conative, this method was problematic because of large differences between the relative frequencies of the alternating SCFs. The data for the rare frame was swamped by the data for the predominant one, resulting in a substantial number of false positives.

There is an implicit threshold when using this method. This threshold is at the combined cost of the separate models. The cost of the model for the combined data has to be *below* this threshold for a verdict of participation. This threshold makes the most sense in MDL terms, however, it is possible that better results might be obtained if a different threshold was used. This could be empirically determined. Setting a threshold somewhere below the cost of the separate models makes the task more stringent, and will reduce the number of false positives.

We proposed a second method in chapter 5 for using PTCMs to detect RSAs. In this method, we compare the probability distributions at the PTCMs of the alternating slots in the alternating frames. Using probability distributions avoids the problem of different relative frequencies of the target SCFs. We established a significant relationship between the similarity of the PTCMs at the target slots and participation using a number of measures of distributional similarity. To obtain figures for precision and recall we used a threshold on the similarity score to determine verbal participation. We experimented with results obtained when using both the mean, and the median, of the scores for the threshold.

As a baseline, we compared the two approaches for identifying verbal participants with TCMs to a baseline approach using a measure of lemma overlap (LO) of the argument head data at the target slots. There was not always a significant relationship between LO and participation. [1] Identification of participants was performed by comparing the LO score to a threshold. As with the experiments using distributional similarity scores and the PTCMs, we obtained the thresholds using means and medians from the sample of positive and negative candidates.

In our results, neither comparing the TCMs at a more specific level than the WordNet roots, nor

[1] In the experiments reported here, we only obtained a significant result with one out of four data sets. In more recent work (McCarthy, 2000) we obtained a significant result in one out of two data sets. The significant result was obtained at a lower significance level than that obtained for the class-based experiments. The lemma-based result was significant at the 95% level, whereas the class-based experiments were significant well above the 99% level.

using WSD to achieve more specific TCMs, significantly improved identification. These findings imply that although it is best to generalise to semantic classes, rather than use the argument head lemmas directly, one can do as well by simply calculating the probability distribution at the WordNet roots, given a particular slot and SCF. It does not appear necessary to seek a more intuitive level of generalisation using MDL. However, it is possible that some verbs might require more specific TCMs. This needs further investigation. Such information could potentially be used in the encoding and consequent description length cost.

In the MDL experiments, we tried ignoring verbs where the semantic type of the arguments was similar when the target slots co-occurred in the same frame. This increased accuracy, at the expense of coverage. For example, this is the case for the causative alternation since the object in the transitive switches to occupy the subject slot in the intransitive. Subject and object slots are both present in the transitive frame. We obtained TCMs at these two slots in the transitive frame and filtered out verbs with similar argument heads across these two slots before applying our approach to the object of the transitive and the subject of the intransitive. Frequently the apparent semantic similarity between co-occurring slots with different grammatical roles in a SCF arose because the differences were within one specific area of WordNet. These differences were not detected because they were overshadowed by similarity at the majority of the classes under the cut. Our approach considered semantic similarity using corpus data in terms of WordNet structure. Automatic classifications might be better placed to highlight semantic similarities and differences apparent in the data.

6.2 Directions for Future Research

In this thesis, we have showed that automatically acquired subcategorization information and selectional preferences can be used to detect role switching alternations. A significant advantage of using probabilistic preference models was that measures of distributional similarity could be used. There was no requirement for a human to specify a priori semantic cues. The most obvious obstacle to automatic acquisition of alternations is the sparseness of the data. This restricted our experiments to a limited set of alternations because we needed a reasonably sized set of positive and negative candidates that took the required SCFs for evaluation purposes. The lack of candidates for many alternations arose because of a combination of the following factors:

- for many alternations, there are only a few verbs which participate
- many alternations involve rare verbs
- for many verbs which participate in an alternation, one of the alternate forms is rare

The first two issues were a problem for evaluation, since we needed a large enough sample of candidates to determine if our method behaved significantly better than a random split. The second issue is also a problem for application, if one wishes to handle rare verbs. The third problem is a problem for both evaluation and application. Problems of sparse data affect identification of alternations because, even with relatively common verbs, one of the alternating variants may not occur with sufficient frequency for selectional preference acquisition.

Our method is generally applicable to RSAs and we predict that it will work with all such alternations, provided that sufficient corpus data is available. Since our approach is fully automatic we do not foresee a problem with obtaining and processing sufficient data.

In addition to increasing the volume of data, there are other directions for future research which might increase coverage, or accuracy, or both.

The SCF acquisition system has a major impact on both coverage and accuracy. We obtained our candidates for a given alternation by virtue of the verbs which were recorded with sufficient entries at the alternating variants in the SCF lexicon. In recent work, Korhonen et al. (2000) demonstrate that the SCF acquisition system shows a good deal of room for improvement due to errors of the statistical filter. Many of the errors are shown to involve medium and low frequency frames. Often, at least one of these frames will be involved in a given alternation. Improving the accuracy of the statistical filter of the SCF acquisition system will shift some FNs, to become TPs, thereby increasing the number of candidates with the appropriate SCF. Shifting FPs to TNs will remove some erroneous candidates, and should increase precision for diathesis alternation detection.

Levin (1993) has shown that groups of diathesis alternations can be used to classify verbs. The alternation behaviour of class members is provided within her verbal classification. Thus, once a verb is placed in the Levin taxonomy, one can predict unseen alternation behaviour. On the other hand, evidence about alternation behaviour can be used to classify a verb within this taxonomy. Research into combining the evidence from different diathesis alternations to classify a verb should help in overcoming the sparse data problem and produce more reliable results. For our approach, this amounts to devising a way of combining the evidence given by the distributional similarity scores for groups of alternations known to be characteristic of a verb class. One possibility is to cluster the similarity scores obtained over the full set of RSAs. Features for other alternations, such selectional preference strength, which has been shown to be an indicator of the implicit object construction (Resnik, 1993a), could also be input to the clustering process.

Many alternations for which we did not have sufficient data for experimentation involve prepositional phrases at the target slots. In the current system, the TCMs are acquired specific to the verb and preposition combination for these slots. It would be possible to back off to preposition classes in cases of sparse data. Preposition classes might be obtained manually, or by clustering prepositions according to the distribution of nominal argument heads in the NPs that they subcategorize. [2] Backing off to preposition classes would reduce the sparse data situation but would also reduce the accuracy of the system, as verbs that alternate do not necessarily do so with all potential prepositions.

There are many further modifications that could be made to the selectional preference acquisition system. For example, one might investigate alternative ways of handling multiple parentage in WordNet. However, from the research presented in chapters 2, 3 and 4, alterations to the selectional preference acquisition system have not significantly affected diathesis alternation detection. There are however three areas that would be worth further investigation:

[2] We performed some preliminary work in this direction using hierarchical clustering. Many of the preposition classes were intuitive, for example *since* and *until* were grouped together. We have not applied this work to diathesis alternation detection. One example where this might work is *put* which subcategorizes for a locative PP. It remains to be seen whether enough other verbs also subcategorize for PPs headed by a preposition class.

1. a more reliable method of disambiguating the argument head data
2. a method of disambiguating the verbs, and differentiating the SCF entries and TCMs accordingly
3. use of an automatically constructed semantic hierarchy

The methods of WSD that we experimented with, handled a large proportion of the argument heads, at the expense of accuracy. The more specific, and more intuitive, preference models did not give rise to a significant improvement in performance for WSD, pseudo-disambiguation or diathesis alternation detection. A reliable method of disambiguating argument heads might, however, be useful in cases of sparse data. One possibility would be to concentrate effort on finding a reliable method for WSD of frequent nouns, which cover a larger portion of the data, rather than attempting to disambiguate all the argument heads. This might achieve better results for verb and slot combinations with little data where, without disambiguation, the TCM is located at the root.

Rudimentary WSD of the argument head data did not improve diathesis alternation identification. One outstanding issue is whether disambiguating the verb forms might help. At the moment SCFs, and therefore selectional preference models, are acquired with respect to a verb form, rather than a verb sense. If one were to disambiguate the verb forms, obtaining an appropriate sense inventory is a complicated matter. One would not want to separate related verb senses which might occur as alternating variants. If an automatic clustering approach is adopted then a promising approach is to consider verb and argument head types together (Rooth et al., 1999), since verbs are reported to be best disambiguated by their argument heads (Manning & Schütze, 1999). Soft clustering permits a verb to belong to more than one class, with a probability distribution associated with each verb over the classes in the classification (Pereira et al., 1993; Rooth et al., 1999). The resulting classification highlights the salient classes (senses) for a verb form. There is a promising aspect of clustering the argument heads alongside the verbs with specified SCF slots, with regard to diathesis alternation identification. A verbs alternation behaviour is brought out by the shared grouping of two alternating entries for a particular verb form, differing in respect of the specified SCF and slot for the argument heads in the cluster Rooth et al. (1999).

In the current framework, reliance on a manmade sense inventory for verbs would not be helpful, since this presupposes knowledge about the items for which we are acquiring information. It makes more sense not to disambiguate the verbs in advance, but to use the Levin classification to disambiguate the verbs as Dorr & Jones (1996) do. If alternations are detected collectively by combining evidence to place the verbs within the Levin taxonomy, then the resulting classification will signify the relevant senses of the verb, even if these were not stipulated in the input data.

One promising avenue for research into diathesis alternation detection is to use an automatically constructed semantic hierarchy for characterising selectional preference models. An automatically constructed hierarchy would be better placed to highlight semantic differences that are apparent in the corpus data. Another advantage is that the method could then be applied to another language, provided the corpus data was available and a shallow parser was available for that language. However, sparse data would still be a problem with selectional preferences represented in automatically constructed hierarchies. It may be even more problematic than when using manu-

ally constructed hierarchies if more tokens of each type are required for classification (Schulte im Walde, 1998).

We have successfully used our method with a threshold of 10 or more argument head instances for each verb and SCF combination. It may be possible to lower this threshold further, particularly in cases where the argument head data falls within the same area of the semantic taxonomy. Certainly, combining evidence from different diathesis alternations should help, provided that there is at least some evidence to start off classification.

Bibliography

Abe, N., & Li, H. (1996). Learning word association norms using tree cut pair models. In *Proceedings of the 13th International Conference on Machine Learning, ICML*, pp. 3–11.

Abney, S., & Light, M. (1999). Hiding a semantic class hierarchy in a Markov model. Technical report WP5.2, SPARKLE Deliverable - IMS, University of Stuttgart.

Agirre, E., & Rigau, G. (1996). Word sense disambiguation using conceptual density. In *Proceedings of the 16th International Conference of Computational Linguistics, COLING-96*, pp. 16–22.

Atkins, S. (1993). Tools for computer-aided lexicography: the Hector project. In *Papers in Computational Lexicography: COMPLEX 93* Budapest.

Basili, R., Pazienza, M. T., & Velardi, P. (1993). Hierarchical clustering of verbs. In Boguraev, B., & Pustejovsky, J. (eds.), *The Acquisition of Lexical Knowledge from Text. SIGLEX ACL Workshop*, pp. 70–81 Columbus Ohio.

Beckwith, R., Fellbaum, C., Gross, D., & Miller, G. A. (1991). WordNet: A lexical database organised on psycholinguistic principles. In Zernik, U. (ed.), *Lexical Acquisition: Exploiting On-Line Resources to Build a Lexicon*, pp. 211–232. Lawrence Erlbaum Associates, Hillsdale NJ.

Boguraev, B., Briscoe, E., Carroll, J., Carter, D., & Grover, C. (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*, pp. 193–200 Stanford, CA.

Boguraev, B., & Briscoe, T. (1987). Large lexicons for natural language processing: utilising the grammar coding system of LDOCE. *Computational Linguistics*, *13*(3-4), 203–218.

Bredenkamp, A., Markantonatou, S., & Sadler, L. (1996). Lexical rules: What are they?. In *Proceedings of the 16th International Conference of Computational Linguistics, COLING-96*, pp. 163–168.

Brent, M. R. (1991). Automatic acquisition of subcategorization frames from untagged text. In *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics*, pp. 209–214.

Brent, M. R. (1993). From grammar to lexicon: unsupervised learning of lexical syntax. *Computational Linguistics*, *19*(2), 243–262.

Briscoe, T., & Carroll, J. (1997). Automatic extraction of subcategorization from corpora. In *Proceedings of the Fifth Applied Natural Language Processing Conference*, pp. 356–363.

Briscoe, T., & Copestake, A. (1996). Controlling the application of lexical rules. In Viegas, E. (ed.), *SIGLEX Workshop on Lexical Semantics - ACL 96 Workshop*, pp. 7–19.

Briscoe, T., & Copestake, A. (1999). Lexical rules in constraint-based grammar. *Computational Linguistics*, *24*(4), 487–526.

Brown, P., Pietra, S., Pietra, V., & Mercer, R. (1991). Word-sense disambiguation using statistical methods. In *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics*, pp. 264–270.

Carroll, G., & Rooth, M. (1998). Valence induction with a head-lexicalized PCFG. In *3rd Conference on Empirical Methods in Natural Language Processing*, pp. 00–00 Granada, Spain.

Carroll, J., & McCarthy, D. (2000). Word sense disambiguation using automatically acquired verbal preferences. *Computers and the Humanities. Senseval Special Issue*, *34*(1–2), 109–114.

Chapman, R. (1977). *Roget's International Thesaurus (Fourth Edition)*. Harper and Row, New York.

Charniak, E. (1993). *Statistical Language Learning*. MIT Press, Cambridge, MA.

Chitrao, M., & Grishman, R. (1990). Statistical parsing of messages. In *DARPA Speech and Natural Language Workshop*, pp. 263–266 Hidden Valley, PA.

Church, K., Gale, W., Hanks, P., & Hindle, D. (1991). Using statistics in lexical analysis. In Zernik, U. (ed.), *Lexical Acquisition. : Exploiting On-Line Resources to Build a Lexicon*, pp. 115–164. Lawrence Erlbaum Associates, Hillsdale NJ.

Clark, S., & Weir, D. (1999). An iterative approach to estimating frequencies over a semantic hierarchy. In *Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, pp. 258–265.

Cover, Thomas, M., & Thomas, J. A. (1991). *Elements of Information Theory*. Wiley-Interscience, New York.

Cowie, J., Guthrie, J. A., & Guthrie, L. (1992). Lexical disambiguation using simulated annealing. In *Proceedings of the 14th International Conference of Computational Linguistics. COLING-92*, Vol. I, pp. 359–365.

Cunningham, H., Gaizauskas, R., & Wilks, Y. (1995). A general architecture for text engineering (GATE) — a new approach to language R&D. Technical report, University of Sheffield, UK, Department of Computer Science.

Dagan, I., Marcus, S., & Markovitch, S. (1993). Contextual word similarity and estimation from sparse data. In *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics*, pp. 164–171.

Dang, H. T., Kipper, K., Palmer, M., & Rosensweig, J. (1998). Investigating regular sense extensions based on intersective Levin classes. In *Proceedings of the 17th International Conference on Computational Linguistics and the 36th Annual Meeting of the Association for Computational Linguistics*, Vol. 1, pp. 293–299.

Dempster, A., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, *39*(B), 1–38.

Dorr, B. J., & Jones, D. (1996). Role of word sense disambiguation in lexical acquisition: predicting semantics from syntactic cues. In *Proceedings of the 16th International Conference on Computational Linguistics, COLING-96*, pp. 322–327.

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

Dunning, T. (1998). Finding structure in text, genome and other symbolic sequences. Unpublished. To obtain a copy contact ted@crl.nmsu.edu.

Elworthy, D. (1994). Does Baum-Welch re-estimation help taggers?. In *Proceedings of the 4th ACL Conference on Applied Natural Language Processing*, pp. 53–58.

Federici, S., Montemagni, S., & Pirrelli, V. (1997). Inferring semantic similarity from distributional evidence: an analogy-based approach to word sense disambiguation. In *Proceedings of the ACL/EACL 97 Workshop Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications*, pp. 90–97.

Federici, S., Montemagni, S., & Pirrelli, V. (1999). Sense: an analogy-based word sense disambiguation system. *Natural Language Engineering*, *5*(2), 207–218.

Federici, S., Montemagni, S., & Pirrelli, V. (2000). ROMANSEVAL: results for Italian by SENSE. *Computers and the Humanities. Senseval Special Issue*, *34*(1–2), 199–204.

Fellbaum, C. (ed.). (1998). *WordNet, An Electronic Lexical Database*. The MIT Press, Cambridge, MA.

Fillmore, C. (1970). The grammar of *hitting* and *breaking*. In Jacobs, R. A., & Rosenbaum, P. S. (eds.), *Readings in English Transformational Grammar*, pp. 120–133. Ginn and Company: a Xerox Company, Waltham M.A.

Finch, S., & Chater, N. (1991). A hybrid approach to the automatic learning of lingusitic categories. *Quarterly Newsletter of the Society for the Study of Artificial Intelligence and Simulation of Behaviour*, *78*, 16–24.

Firth, J. R. (1957). A synopsis of linguistic theory 1930-1955. In *Studies in Linguistic Analysis*. Blackwell, Oxford. Reprinted in F.R. Palmer. (ed.), *Selected Papers of J.R. Firth*. Longman. 1968. pp.168–205.

Francis, W., & Kučera, H. (1979). *Manual of Information to Accompany a Standard Corpus of Present-Day Edited American English, for Use with Digital Computers*. Department of Linguistics, Brown University, Rhode Island. Revised and amplified ed.

Gaizauskas, R., & Humphreys, K. (1996). Using verb semantic role information to extend partial parses via a co-reference mechanism. In Carroll, J. (ed.), *Proceedings of the 8th European Summer School in Logic, Language and Information ESSLLI96 - Workshop on Robust Parsing*, pp. 103–113.

Gale, W., Church, K., & Yarowsky, D. (1992). Estimating upper and lower bounds on the performance of word-sense disambiguation programs. In *Proceedings of the 30th Annual Meeting of the Association for Computational Linguistics*, pp. 249–256.

Garside, R., Leech, G., & Sampson, G. (1987). *The Computational Analysis of English: A Corpus-Based Approach*. Longman, London.

Gazdar, G. (1996). Paradigm merger in natural language processing. In Milner, R., & Wand, I. (eds.), *Computing Tomorrow: Future Research Directions in Computer Science*, pp. 88–109. Cambridge University Press, Cambridge, UK.

Ge, N., Hale, J., & Charniak, E. (1998). A statistical approach to anaphora resolution. In *Proceedings of the Sixth Workshop on Very Large Corpora*, pp. 161–170. Montreal, Canada. Association of Computational Linguistics.

Grishman, R., Macleod, C., & Meyers, A. (1994). Comlex syntax: building a computational lexicon. In *Proceedings of the 15th International Conference of Computational Linguistics, COLING-94*, pp. 268–272 Kyoto, Japan.

Grishman, R., & Sterling, J. (1993). Smoothing of automatically generated selectional constraints. In *Proceedings of the ARPA Workshop on Human Language Technology*, pp. 254–259. Morgan Kaufman.

Hanks, P. (ed.). (1979). *Collins English Dictionary*. Collins, London & Glasgow.

Hindle, D., & Rooth, M. (1991). Structural ambiguity and lexical relations. In *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics*, pp. 229–236.

Hindle, D., & Rooth, M. (1993). Structural ambiguity and lexical relations. *Computational Linguistics*, *19*(1), 103–120.

Hornby, A. S. (1989). *Oxford Advanced Learner's Dictionary of Current English*. Oxford University Press, Oxford.

Howitt, D., & Cramer, D. (1997). *An Introduction to Statistics for Psychology: a Complete Guide for Students*. Prentice Hall, London.

Ide, N., & Véronis, J. (1998). Introduction to the special issue on word sense disambiguation: the state of the art. *Computational Linguistics*, *24*(1), 1–40.

Inui, K., Sornlertlamvanich, V., Tanaka, H., & Tokunaga, T. (1997). A new formalization of probabilistic GLR parsing. In *5th ACL/SIGPARSE International Workshop on Parsing Technologies*, pp. 123–134 Cambridge, MA.

Jackendoff, R. (1990). *Semantic Structures*. MIT Press, Cambridge, MA.

Johnansson, S., Leech, G., & Goodluck, H. (1978). *Manual of Information to Accompany the Lancaster-Oslo/Bergen Corpus of British English, for Use with Digital Computers*. Department of English, University of Oslo, Oslo.

Katz, J., & Fodor, J. (1964). The structure of a semantic theory. In Fodor, J., & Katz, J. (eds.), *The Structure of Language*, chap. 19, pp. 479–518. Prentice Hall.

Kilgarriff, A. (1993). Dictionary word-sense distinctions: an enquiry into their nature. *Computers and the Humanities*, *26*(1–2), 365–387.

Kilgarriff, A., et al. (1998). SENSEVAL - evaluating word sense disambiguation systems. http://www.itri.brighton.ac.uk/events/senseval/proceedings.

Kilgarriff, A., & Palmer, M. (eds.). (2000). *Senseval: Special Issue of the Journal Computers and the Humanities*, Vol. 34(1–2). Kluwer, Dordrecht, the Netherlands.

Kilgarriff, A., & Rosenzweig, J. (2000). Framework and results for english SENSEVAL. *Computers and the Humanities. Senseval Special Issue*, *34*(1–2), 15–48.

Klavans, J. L., & Resnik, P. (eds.). (1996). *The Balancing Act*. The MIT Press, Cambridge, MA.

Knight, K., & Luk, S. (1994). Building a large knowledge base for MT. In *Proceedings of the Twelfth Conference of the American Association for Artificial Intelligence - AAAI*, pp. 185–209 Seattle,WA.

Kohl, K., Jones, D., Berwick, R., & Nomura, N. (1998). Representing verb alternations in WordNet. In Fellbaum, C. (ed.), *WordNet: an Electronic Lexical Database*. MIT Press, Cambridge, MA.

Korhonen, A. (1997). *Acquiring Subcategorisation from Textual Corpora*. Master's thesis, University of Cambridge.

Korhonen, A. (1998). Automatic extraction of subcategorization frames from corpora - improving filtering with diathesis alternations. In *Proceedings of the ESSLLI 98 Workshop on Automated Acquisition of Syntax and Parsing*, pp. 49–56 Saarbrucken, Germany.

Korhonen, A., Gorrell, G., & McCarthy, D. (2000). Statistical filtering and subcategorization frame acquisition. In *Proceedings of the 2000 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora.*, pp. 199–206 Hong Kong. ACL.

Krenn, B., & Samuelsson, C. (1997). *The Linguist's Guide to Statistics*. http://coli.uni-sb.de/ krenn.

Lapata, M. (1999). Acquiring lexical generalizations from corpora: a case study for diathesis alternations. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, pp. 397–404.

Leacock, C., Towell, G., & Voorhees, E. (1993). Corpus-based statistical sense resolution. In *Proceedings of the ARPA Workshop on Human Language Technology*, pp. 260–265. Morgan Kaufman.

Lee, L. (1997). *Similarity-Based Approaches to Natural Language Processing*. Ph.D. thesis, Harvard.

Lee, L. (1999). Measures of distributional similarity. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, pp. 25–32.

Leech, G. (1992). 100 million words of English: the British National Corpus. *Language Research*, *28*(1), 1–13.

Lesk, M. (1986). Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from and ice cream cone. In *Proceedings of the ACM SIGDOC Conference*, pp. 24–26 Toronto, Canada.

Levin, B. (1993). *English Verb Classes and Alternations: a Preliminary Investigation*. University of Chicago Press, Chicago and London.

Levin, B., & Rappaport Hovav, M. (1995). *Unaccusativity: at the Syntax-Lexical Semantics Interface*. MIT Press, Cambridge, MA.

Levin, B., & Rappaport Hovav, M. (1996). Lexical semantics and syntactic structure. In Lappin, S. (ed.), *The Handbook of Contemporary Semantic Theory*. Blackwell, Cambridge MA.

Li, H. (1998). *A Probabilistic Approach to Lexical Semantic Knowledge Acquisition and Structural Disambiguation*. Ph.D. thesis, University of Tokyo.

Li, H., & Abe, N. (1995). Generalizing case frames using a thesaurus and the MDL principle. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing*, pp. 239–248 Bulgaria.

Li, H., & Abe, N. (1996). Clustering words with the MDL principle. In *Proceedings of the 16th International Conference of Computational Linguistics. COLING-96*, pp. 4–9.

Li, H., & Abe, N. (1998). Generalizing case frames using a thesaurus and the MDL principle. *Computational Linguistics*, *24*(2), 217–244.

Lyons, J. (1977). *Semantics*, Vol. 2. CUP, Cambridge.

Manning, C. (1993). Automatic acquisition of a large subcategorization dictionary from corpora. In *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics*, pp. 235–242.

Manning, C., & Schütze, H. (1999). *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, MA.

Marcus, M., et al. (1995). The Penn Treebank: annotating predicate argument structure. Technical report, University of Pennsylvania. Distributed on The Penn Treebank 2 CD-ROM by the Linguistic Data Consortium.

McCarthur, T. (ed.). (1981). *Longman Lexicon of Contemporary English*. Longman Group Ltd., London, UK.

McCarthy, D. (1997). Word sense disambiguation for acquisition of selectional preferences. In *Proceedings of the ACL/EACL 97 Workshop Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications*, pp. 52–61.

McCarthy, D. (2000). Using semantic preferences to identify verbal participation in role switching alternations. In *Proceedings of the First Conference of the North American Chapter of the Association for Computational Linguistics. (NAACL)*, pp. 256–263 Seattle,WA.

McCarthy, D., & Korhonen, A. (1998). Detecting verbal participation in diathesis alternations. In *Proceedings of the 17th International Conference on Computational Linguistics and the 36th Annual Meeting of the Association for Computational Linguistics*, Vol. 2, pp. 1493–1495.

McKeown, K., & Hatzivassiloglou, V. (1993a). Augmenting lexicons automatically: clustering semantically related adjectives. In *Proceedings of the ARPA Workshop on Human Language Technology*, pp. 272–277. Morgan Kaufman.

McKeown, K., & Hatzivassiloglou, V. (1993b). Towards the automatic identification of adjectival scales: clustering adjectives according to meaning. In *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics*, pp. 172–182.

Miller, George, A., Leacock, C., Tengi, R., & Bunker, R. T. (1993a). A semantic concordance. In *Proceedings of the ARPA Workshop on Human Language Technology*, pp. 303–308. Morgan Kaufman.

Miller, G., Beckwith, R., Fellbaum, C., Gross, D., & Miller, K. (1993b). *Introduction to WordNet: an On-Line Lexical Database*. ftp://clarity.princeton.edu/pub/WordNet/5papers.ps.

Montemagni, S. (1994). Extracting typical subjects and objects of verbs from mono- and bi-lingual dictionaries. Technical report, ACQUILEX-II. http://www.cl.cam.ac.uk/Research/NL/acquilex/acq2wps.html.

Montemagni, S., & Pirrelli, V. (1995). Do lexical rules apply across the board? a corpus-based investigation in the machinery of the causative-inchoative alternation in Italian. Technical report 71, ACQUILEX-II.

Montemagni, S., Pirrelli, V., & Ruimy, N. (1995). Ringing things which nobody can ring: a corpus-based study of the causative-inchoative alternation in Italian. *Textus*, *VIII*, 371–390.

MUC-4 (1992). *Proceedings of the Fourth Message Understanding Conference*. Morgan Kaufmann, San Mateo, CA.

Murata, M., Isahara, H., & Nagao, M. (1999). Resolution of indirect anaphora in Japanese sentences using examples 'x no y (y of x)'. In *Proceedings of the ACL'99 Workshop on 'Coreference and Its Applications'* Maryland, USA.

Ng, H. T., & Lee, H. B. (1996). Integrating multiple knowledge sources to disambiguate word sense: an exemplar-based approach. In *Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics*, pp. 40–47.

Nicholls, D. (1994). French and English diathesis alternations and the LKB. Technical report 44, ACQUILEX-II.

Nicholls, D. (1995). Can fully-productive lexical rules be defined and can they be defined cross linguistically? A corpus-based study of English and French verb argument structures. Technical report 79, ACQUILEX-II.

Pedersen, T. (1996). Fishing for exactness. In *Proceedings of the South-Central SAS Users Group Conference SCSUG-96*, pp. 00–00 Austin, Texas.

Pereira, F., Tishby, N., & Lee, L. (1993). Distributional clustering of English words. In *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics*, pp. 183–190.

Pinker, S. (1989). *Learnability and Cognition*. MIT Press, Cambridge MA.

Pirrelli, V., Ruimy, N., & Montemagni, S. (1994). Lexical regularities and lexicon compilation. a case study: argument structure alternations of Italian verbs. Technical report 36, ACQUILEX-II.

Pollard, C., & Sag, I. (1987). *An Information-Based Approach to Syntax and Semantics: Volume 1 Fundamentals*. CSLI Lecture Notes 13, Stanford CA.

Pollard, C., & Sag, I. (1994). *Head-Driven Phrase Structure Grammar*. Chicago University Press, Chicago.

Pozanski, V., & Sanfilippo, A. (1996). Detecting dependencies between semantic verb subclasses and subcategorization frames in text corpora. In Boguraev, B., & Pustejovsky, J. (eds.), *Corpus Processing and Lexical Acquisition*, pp. 175–190. MIT Press, London, England.

Price, P. (1996). Combining linguistic with statistical methodology in automatic speech understanding. In Klavans, J. L., & Resnik, P. (eds.), *The Balancing Act*. The MIT Press, Cambridge, MA.

Procter, P. (ed.). (1978). *Longman Dictionary of Contemporary English*. Longman Group Ltd., Harlow, UK.

Procter, P. (ed.). (1995). *Cambridge International Dictionary of English*. Cambridge University Press, Cambridge, UK.

Radford, A. (1989). *Transformational Grammar*. Cambridge University Press, Cambridge, UK.

Resnik, P. (1992). A class-based approach to lexical discovery. In *Proceedings of the 30th Annual Meeting of the Association for Computational Linguistics*, pp. 327–329.

Resnik, P. (1993a). *Selection and Information: A Class-Based Approach to Lexical Relationships*. Ph.D. thesis, University of Pennsylvania.

Resnik, P. (1993b). Semantic classes and syntactic ambiguity. In *Proceedings of the ARPA Workshop on Human Language Technology*, pp. 278–283. Morgan Kaufman.

Resnik, P. (1997). Selectional preference and sense disambiguation. In *Proceedings of the* SIGLEX *Workshop on Tagging Text with Lexical Semantics: Why What and How?*, pp. 52–57 Washington, DC.

Resnik, P., & Yarowsky, D. (1997). A perspective on word sense disambiguation methods and their evaluation. In *Proceedings of the* SIGLEX *Workshop on Tagging Text with Lexical Semantics: Why, What and How?*, pp. 79–86 Washington, DC.

Ribas, F. (1994). Learning more appropriate selectional restrictions. Technical report 41, ACQUILEX-II.

Ribas, F. (1995a). *On Acquiring Appropriate Selectional Restrictions from Corpora Using a Semantic Taxonomy*. Ph.D. thesis, University of Catalonia.

Ribas, F. (1995b). On learning more appropriate selectional restrictions. In *Proceedings of the Seventh Conference of the European Chapter of the Association for Computational Linguistics*, pp. 112–118.

Rissanen, J. (1978). Modeling by shortest data description. *Automatica*, *14*, 465–471.

Rissanen, J. (1986). Stochastic complexity and modeling. *The Annals of Statistics*, *14*(3), 1080–1100.

Rooth, M. (1998). Two-dimensional clusters in grammatical relations. In *Inducing Lexicons with the* EM *algorithm*, Vol. Aims Report 4(3), pp. 7–24. IMS, University of Stuttgart.

Rooth, M., Riezler, S., Prescher, D., Carroll, G., & Beil, F. (1999). Inducing a semantically annotated lexicon via EM-based clustering. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, pp. 104–111.

Rosenzweig, J. (1998). SENSEVAL scores. http://www.itri.brighton.ac.uk/seminarindex.html.

Sampson, G. (1995). *English for the Computer*. Oxford University Press.

Sampson, G. (2000). Book review: "wordnet", ed. christiane fellbaum. *International Journal of Lexicography*, *13*, 54–59.

Sanfilippo, A. (1994). Word knowledge acquisition, lexicon construction and dictionary compilation. In *Proceedings of the 15th International Conference of Computational Linguistics. COLING-94*, Vol. I, pp. 273–277.

Sanfilippo, A. (1996). LKB encoding of lexical knowledge from machine readable dictionaries. In Briscoe, E., Copestake, A., & de Paiva, V. (eds.), *Inheritance, Defaults and the Lexicon*, pp. 190–222. Cambridge University Press, Cambridge.

Schulte im Walde, S. (1998). Automatic semantic classification of verbs according to their alternations. In *Inducing Lexicons with the* EM *algorithm*, Vol. Aims Report 4(3), pp. 55–74. IMS, University of Stuttgart.

Schütze, H. (1992). Dimensions of meaning. In *Supercomputing*, pp. 787–796 Minneapolis.

Schütze, H. (1993). Part-of-speech induction from scratch. In *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics*, pp. 251–258.

Schütze, H. (1998). Automatic word sense discrimination. *Computational Linguistics*, *24*(1), 97–123.

Schütze, H., & Pederson, J. O. (1995). Information retrieval based on word senses. In *Proceedings of the Fourth Annual Symposium on Document Analysis and Information Retrieval*, pp. 161–175 Las Vegas, NV.

Sekine, S., Ananiadou, S., Carroll, J., & Tsuji, J. (1992). Linguistic knowledge generator. In *Proceedings of the 14th International Conference of Computational Linguistics. COLING-92*, pp. 560–566.

Siegel, S., & Castellan, N. J. (eds.). (1988). *Non-Parametric Statistics for the Behavioral Sciences*. McGraw-Hill, New York.

Sinclair, J. (ed.). (1987). *Collins COBUILD English Language Dictionary*. Collins, London.

Slator, B., & Wilks, Y. (1990). Towards semantic structures from dictionary entries. In Schmitz, U., Schuetz, R., & Kunz, A. (eds.), *Linguistic Approaches to Artificial Intelligence*, pp. 419–460. Peter Lang, Frankfurt.

Smadja, F. (1993). Retrieving collocations from text: Xtract. *Computational Linguistics. Special Issue on Using Large Corpora*, *19*(1), 143.

Sparck Jones, K., & Galliers, J. R. (1996). *Evaluating Natural Language Processing Systems: An Analysis and Review*. Springer, London.

Stede, M. (1998). A generative perspective on verb alternations. *Computational Linguistics*, *24*(3), 401–430.

Stevenson, R. M. (1999). *Multiple Knowledge Sources for Word Sense Disambiguation*. Ph.D. thesis, University of Sheffield.

Stevenson, S., & Merlo, P. (1999). Automatic verb classification using distributions of grammatical features. In *Proceedings of the Nineth Conference of the European Chapter of the Association for Computational Linguistics*, pp. 45–52.

Ushioda, A., Evans, D., Gibson, T., & Waibel, A. (1993). The automatic acquisition of frequencies of verb subcategorization frames from tagged corpora. In Boguraev, B., & Pustejovsky, J. (eds.), *The Acquisition of Lexical Knowledge from Text. SIGLEX ACL Workshop*, pp. 95–106 Columbus Ohio.

Veronis, J., & Ide, N. (1990). Word sense disambiguation with very large neural networks extracted from machine readable dictionaries. In *Proceedings of the 13th International Conference of Computational Linguistics. COLING-90*, Vol. II, pp. 389–394.

Vossen, P. (1999). EuroWordNet general document. Technical report, University of Amsterdam. http://www.hum.uva.nl/ ewn/.

Wagner, A. (2000). Enriching a lexical semantic net with selectional preferences by means of statistical corpus analysis. In *Proceedings of the ECAI-2000 Workshop on Ontology Learning*, pp. 00–00 Berlin.

Wakao, T., Gaizauskas, R., & Wilks, Y. (1996). Evaluation of an algorithm for the recognition and classification of proper names. In *Proceedings of the 16th International Conference of Computational Linguistics. COLING-96*, pp. 418–423.

Wilks, Y. (1975a). An intelligent analyzer and understander of English. *CACM*, *18*(5), 264–274. Reprinted in Grosz,B. & Sparck Jones, K. & Webber,B. (eds.), *Readings in Natural Language Processing*. Morgan Kaufmann. 1986. pp.193–203.

Wilks, Y. (1975b). Preference semantics. In Keenan, Edward, L. (ed.), *Formal Semantics of Natural Language*, pp. 329–348. Cambridge University Press. Papers from the Colloquium on Formal Semantics of Natural Language (1973) sponsored by the King's College Research Centre, Cambridge.

Wilks, Y., & Stevenson, M. (1998a). The grammar of sense: using part-of speech tags as a first step in semantic disambiguation. *Natural Language Engineering*, *4*(2), 135–143.

Wilks, Y., & Stevenson, M. (1998b). Optimising combinations of knowledge sources for word sense disambiguation. In *Proceedings of the 17th International Conference on Computational Linguistics and the 36th Annual Meeting of the Association for Computational Linguistics*, pp. 1398–1402 Montreal, Canada.

Yarowsky, D. (1992). Word sense disambiguation using statistical models of Roget's categories trained on large corpora. In *Proceedings of the 14th International Conference on Computational Linguistics. COLING-92*, Vol. II, pp. 454–460.

Yarowsky, D. (1993). One sense per collocation. In *Proceedings of the ARPA Workshop on Human Language Technology*, pp. 266–271. Morgan Kaufman.

Yarowsky, D. (1994). Decision lists for lexical ambiguity resolution: application to accent restoration in Spanish and French. In *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics*, pp. 88–95.

Yarowsky, D. (1995). Unsupervised word sense disambiguation rivaling supervised methods. In *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics*, pp. 189–196.

Zipf, G. K. (1935). *The Psycho-Biology of Language: an Introduction to Dynamic Biology*. MIT Press, Cambridge MA.