LEARNING THEMATIC ROLE RELATIONS FOR LEXICAL SEMANTIC NETS

von

ANDREAS WAGNER

Philosophische Dissertation
angenommen von der Neuphilologischen Fakultät
der Universität Tübingen
am 8. Dezember 2004

Tübingen
2005

Gedruckt mit Genehmigung der Neuphilologischen Fakultät
der Universität Tübingen

Hauptberichterstatter:	Prof. Dr. Erhard Hinrichs
Mitberichterstatter:	Prof. Dr. Martin Volk
Dekan:	Prof. Dr. Joachim Knape

Acknowledgements

This work would not have been possible without the intellectual and/or moral support of a number of persons.

First of all, I wish thank my main supervisor Prof. Erhard Hinrichs and my second supervisor Prof. Martin Volk. Erhard Hinrichs was the person who made this thesis possible by accepting me as a PhD student in the Graduiertenkolleg "Integriertes Linguistik-Studium" at the University of Tübingen. His supervision was characterised by a spirit of great open-mindedness and benevolence, especially during demanding periods of my work. Martin Volk kindly agreed to take on the job of being second supervisor, at a point where my research was already at quite an advanced state. He was thus faced with the difficulty of becoming familiar with a sophisticated system of interdependent data types, processing approaches, and motivations behind these. The necessity of remote communication between Tübingen and Stockholm introduced a further challenge. Consequently, Martin gave invaluable hints from an "outside perspective". Both supervisors provided comprehensive comments on the first drafts of the individual chapters, which helped to enhance the readability and comprehensibility of the final text. Last but not least, they managed to write their official reports about this thesis within the shortest time conceivable so that its defence could take place as soon as five weeks after its submission.

Apart from my supervisors, numerous other people inside and outside of Tübingen have supported me by discussing my research, giving feedback on my talks, or providing technical aid. I am grateful to Steve Abney and Marc Light, who worked here when I started my work. Steve used to thoroughly read and comment on my exposés which comprised my initial efforts to delimit the appropriate aim and scope of the thesis. From him, I have learned the fundamental terms and techniques of statistical NLP. Steve and Marc provided me with the program code they had implemented and the data they used for selectional preference acquisition within their research project. In this way, I could adopt those parts of their implementation which were independent of the actual acquisition method. This substantially facilitated the setup of my early experiments. In the meantime, Steve has made his CASS software package publicly available, which also contains a module for efficiently collecting and processing co-occurrence statistics. I have employed this module for all experiments related to this thesis. I also thank the members of the Stuttgart/Tübingen reading group on statistical NLP, namely Stefan Riezler, Glenn Carroll, Marc Light, Detlef Prescher, and Helmut Schmid. This group was an invaluable forum for improving our expertise and also each discussing our own work. More general discussions, which helped me to think out of the box and locate my research within general linguistics, took place in the Graduiertenkolleg. In particular, I would like to thank Petra Gretsch, Laura Kallmeyer, Anke Lüdeling, and Doris Stolberg for these discussions and, more importantly, for their friendship.

After my time in the Graduiertenkolleg, I joined the projects EuroWordNet-II and GermaNet at the Seminar für Sprachwissenschaft in Tübingen. This environment significantly broadened my perspective on resources like WordNet and EuroWordNet, which were essential to finally fix the scope and the limits of this thesis. Special thanks go to Claudia Kunze, Karin Naumann, and Lothar Lemnitzer for valuable discussions about wordnets and lexical acquisition in general. Moreover, the participation in EuroWordNet was crucial for me to have unrestricted access to the EWN database, which might well have have been too costly otherwise.

When EuroWordNet-II was about to come to an end, I joined the Collaborative Research Centre "Linguistic Data Structures" (SFB 441) in Tübingen, where my duties and responsibilities covered very interesting aspects of corpus linguistics, which, however, hardly overlapped with the work of this thesis. I nevertheless enjoyed the inspiring and stimulating atmosphere and, in particular, the

solidarity of the numerous fellow sufferers who worked on their own PhD theses. Among these, I would like to give special thanks to Sandra Kübler, who was always ready to discuss my ideas and to listen to my professional and personal complaints. Moreover, she proof-read large parts of the final text. In general, I am grateful to all my colleagues for their patience and moral support during periods when I was completely occupied by my PhD project. This especially holds for Dirk Wiebel and Reiner Link, who came to my aid by taking over tasks of mine (primarily related to system administration) in such phases, and to my boss Prof. Marga Reis, who persistently urged me (as well as all other staff members in the same situation) not to neglect PhD research despite our everyday duties.

I am very grateful to a number of people outside of Tübingen for valuable discussions and feedback, which greatly helped me to develop and clarify my ideas. In the first place, I have to mention Diana McCarthy here. We had the opportunity to discuss our PhD projects in great detail and very productively. Our research interests were so similar that we were able to exchange our ideas at a very fine-grained level, yet our ultimate goals were so different that there was no danger of counterproductive rivalry. For me, this cooperation was a stroke of luck. I thank Gerald Gazdar for arranging the contact with Diana. With Philip Resnik, I had an interesting email exchange concerning the general task of this thesis. I also thank Sabine Schulte im Walde for providing me with the training data which I used for my detailed evaluation experiments.

Regarding financial support, I thank the Deutsche Forschungsgemeinschaft (DFG) and the European Union. Without the funding granted by these institutions, it would not have been possible to accomplish this work.

Special thanks go to the people from my circle of friends and relatives who have intensively accompanied my PhD project, sharing my pleasures and sufferings and reminding me that there are still other things in the world than corpora and statistical NLP. In this regard, I would like to emphasise my Tübingen friends, especially Friedhelm Panteleit, Christian Höppler, Wolfgang Huber, Diana Marquardt, Susanne Neuhäusler, Norbert Tausch, and Esther Wedeniwski, who were co-singers of mine in the KHG choir, and, most notably, my parents and my sister Martina. Finally, I am particularly grateful to the most important person in my life, Regina. Although we did not meet until the latest stages of this thesis, her love, cheerfulness, and patience provided a crucial contribution to my eventually finishing it.

Summary

This thesis presents a strategy for the acquisition of thematic role relations (such as AGENT, PATIENT, or INSTRUMENT) by means of statistical corpus analysis, for the purpose of semi-automatically extending lexical-semantic nets. In particular, this work focuses on resources in the style of WordNet (Fellbaum 1998) and EuroWordNet (Vossen 1999). Lexical-semantic nets represent the meanings of words via semantic relations between words and/or word concepts. Semantic (thematic) role relations are conceptual relations which hold between verbs and their nominal arguments (e.g. <eat>–AGENT–<human> or <eat>–PATIENT–<food>). Such relations capture selectional restrictions of verbs. Therefore, the task of acquiring thematic role relations is intrinsically related to the task of acquiring selectional restrictions.

Consequently, the core of a strategy for learning role relations consists in a method for learning selectional restrictions (or, more precisely, selectional preferences). For the latter task, a number of methods have been proposed which utilise syntactically analysed corpora and WordNet. To acquire the selectional preferences of a certain verb for a certain argument, the respective complement nouns of that verb are extracted from the corpus, and statistical methods are applied to generalise over these nouns; these generalisations are expressed as a set of WordNet noun concepts. One of these approaches, namely the method proposed by (Abe & Li 1996), constitutes the starting point of my research. However, this approach is not immediately applicable for learning role relations, but requires modifications and extensions for that task. In particular, two aspects have to be taken into account. Firstly, it is crucial that the WordNet concepts acquired to represent selectional preferences of a verb are located at an appropriate level of generalisation (e.g. <food> as PATIENT of <eat>, rather than <cake> or <physical_object>). I develop a modification of the approach which substantially improves its performance in this respect. Secondly, as the existing methods generalise over syntactic complements, they acquire selectional preferences for syntactic rather than semantic arguments. To learn selectional preferences for semantic roles, the syntactic arguments provided by the parsed corpus have to be linked to their underlying roles so that the statistical learning method can be applied to generalise, for example, over all (semantic) Agents of the examined verb rather than over all its (syntactic) subjects. Therefore, I develop a method for linking syntactic to semantic arguments. A further aspect of the overall strategy I present is an appropriate method for mapping the verbs and nouns in the training data to the corresponding WordNet concepts, which is a prerequisite for applying the preference acquisition algorithm.

To evaluate the role acquisition approach developed in this thesis, I extract a gold standard from the EuroWordNet database and propose detailed evaluation criteria. Overall, the evaluation results (accuracy rates of up to 84%) show that the approach works effectively.

Contents

Chapter 1

Introduction

This thesis deals with the acquisition of thematic role relations by means of statistical corpus analysis, for the purpose of semi-automatically extending lexical-semantic nets. Lexical-semantic nets are resources which represent the semantics of words via semantic relations between words and word concepts. Role relations are semantic relations which hold between verbs and their nominal arguments. Such relations capture selectional restrictions of verbs. Therefore, the task of acquiring thematic role relations is intrinsically related to the task of acquiring selectional restrictions.

This chapter motivates and specifies the goal of the work described in this Ph.D. thesis. Section 1.1 provides a short and informal introduction to selectional restrictions and their importance for natural language processing. Section 1.2 introduces WordNet and EuroWordNet, the lexical-semantic resources which will be used in this work. Section 1.3 defines and delimits the task of this thesis. Section 1.4 provides an overview of the remaining chapters.

1.1 Selectional Restrictions in Natural Language Processing

1.1.1 Motivation

Selectional restrictions (also referred to as *selectional constraints* or *sortal constraints*) can be defined as semantic constraints which a predicate imposes on its arguments. Such constraints account for the anomaly of semantically deviant sentences like

(1.1) *The stone is thinking.

The verb "think" selects a subject which is human; thus, inanimate subjects are rejected by this constraint. Selectional restrictions express the conditions of semantic compatibility between a predicate and its arguments: "think" is not compatible with a non-human subject like "stone".

Several syntactic configurations, i.e. syntactic relations between words and constituents in a sentence, involve selectional restrictions. Sentence (1.1) exemplifies constraints that a verb implies for its complements. In general, the investigation of selectional restrictions has focused on this type of syntactic

dependency. However, other parts-of-speech may represent predicates as well, which also pose semantic constraints on their arguments. For instance, adjectives may semantically restrict the range of the nouns they modify, e.g. “brave” qualifies (groups of) human beings.

Knowledge about selectional restrictions is very useful for basic natural language processing tasks, such as lexical and structural disambiguation or anaphora resolution. In the sentence

(1.2) The dishwasher read the newspaper.

the lexical ambiguity of “dishwasher” (which could denote a human being or a device) is resolved by the selectional restrictions of “read”, which require that the actor expressed by the subject must be human and thus ruling out the alternative ‘device’ sense. The sentence

(1.3) Peter watched the woman with the umbrella.

contains a potential PP-attachment ambiguity: the PP “with the umbrella” could be a modifier of the object noun or a complement of the verb, in which case it would denote an instrument. This structural ambiguity can be resolved by the selectional restrictions of “watch”: these restrictions state that the instrument has to be some optical device, which rules out the possibility of attaching the PP to the verb. In

(1.4) Claire enjoyed the view of the cows on top of the hills, which were mooing gently.

the selectional restrictions of “moo” imply that the relative pronoun “which” refers to “the cows” rather than “the hills”.

The information provided by selectional restrictions alone is generally not sufficient to perform these basic disambiguation tasks. For instance, based on this kind of information it is not possible to resolve the PP-attachment ambiguity in the well-known example

(1.5) Peter watched the man with the telescope.

Nonetheless, examples (1.2)–(1.4) illustrate that such information can make a significant contribution to ambiguity resolution. The practical usefulness of selectional restrictions in this respect has been shown in numerous studies, e.g. (Resnik 1997), (McCarthy, Carroll & Preiss 2001) for word sense disambiguation, (Resnik 1993), (Li 1996) for PP-attachment disambiguation, or (Hobbs 1978) for anaphora resolution.

Resolving ambiguities is a prerequisite for a wide range of elaborate NLP tasks. Therefore, it is not surprising that the application of selectional restrictions has been widely investigated in NLP. Their benefit has been shown for various tasks including speech recognition (cf. (Ueberla 1994)), parsing (cf. (Huyck 2000)), machine translation (cf. (Viegas 1999)), semantic inferencing (cf. (Burns & Davis 1999)), or natural language generation (cf. (Kozlowski, McCoy & Vijay-Shanker 2002)).

1.1.2 Selectional Restrictions in NLP Systems

Information about selectional restrictions has been integrated by some means or other into various NLP systems in order to support disambiguation tasks as sketched in the previous subsection. (Wagner 1995) and (Wagner & Mastropietro 1996) provide an exhaustive discussion of some of these systems, comprising their overall architecture, the individual components and their interaction, and, in particular, the role of selectional restrictions. In this section, I will concentrate on peculiar aspects of encoding selectional restrictions in such systems. These aspects provide valuable insights regarding the work described in this thesis. I will exemplarily refer to three concrete systems.

One of the earliest well-known NLP systems which makes use of selectional restrictions is the *preference semantics* system developed in the seventies by Yorick Wilks (cf. (Wilks 1986)). This is a prototype machine translation (MT) system that determines a semantic representation of a short piece of English text and generates a French text from this representation. One important component of the system is a lexicon comprising about 600 words. In this lexicon, the meaning of a word is represented by a semantic formula, which consists of a binary structure over semantic primitives. (About 70 primitives are defined.) Selectional restrictions are represented as components of such a formula. For example, the formula for "interrogate" is as follows:

(1.6) ((MAN SUBJ)((MAN OBJE)(TELL FORCE)))

This formula says that "interrogate" selects a human subject (MAN SUBJ)[1] and a human object (MAN OBJE) and that the subject forces the object to tell something (TELL FORCE). The first two parts specify selectional restrictions on the verb's arguments: subject and object have to be human.[2] The formulae representing the senses of nouns specify coarse semantic types via the same inventory of primitives that are employed to specify selectional constraints of verbs, e.g. MAN, THING, PART, or FOLK. For example, the ambiguous noun "crook" receives two formulae, one headed by MAN and one headed by THING. In the sentence

(1.7) The policeman interrogates the crook.

the selectional restrictions of "interrogate" disambiguate "crook", since only the MAN sense is compatible with them.

The term *preference semantics* refers to a fundamental design principle of Wilks' system. Semantic definitions, such as selectional constraints, are not treated as strict rules, but as preferences. A sentence which violates such preferences nonetheless receives a semantic interpretation. In case of competing analyses for an utterance, that analysis with the least number of preference violations is selected. This approach accounts for the fact that violations of selectional restrictions are all but exceptional. For example, such violations occur in common rhetorical figures, e.g. in metaphoric expressions like

(1.8) The car drinks gasoline.

[1] In Wilks' notation, MAN represents any human being, not just male humans.

[2] It is important to note that the terms *subject* and *object* in Wilks' representation do not refer to surface syntactic relations, but to semantic deep cases (cf. section 2.1.1).

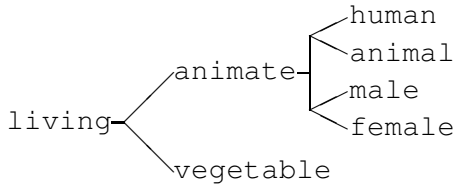


Figure 1.1: Part of the sortal hierarchy in the CLE

or in metonymic constructions like

(1.9) Paris and Berlin have agreed on matters for the French presidency of the EU.

As a consequence of this fact, selectional restrictions are regarded as preferences in this thesis. In section 2.3.2, I will address this issue in more detail. As we will see in chapter 3, the view of selectional restrictions as a preferential phenomenon has generally been adopted within statistical NLP. Essentially, this idea traces back to Wilks' MT system.

The Core Language Engine (CLE) (cf. (Alshawi 1992)) was developed in the late eighties. The CLE was designed as a multi-purpose NLP system, comprising various language processing modules (ranging from spelling correction and morphological analysis to pragmatic plausibility checks) and intended as the core engine for different applications (e.g. natural language interfaces or MT). One module is a lexicon which comprises about 1600 words, being extensible and modifiable for specific applications. Selectional restrictions are encoded and processed in an analogous way to Wilks' system and many other systems. However, there is one striking difference between Wilks' system and the CLE. Wilks organises the inventory of semantic primitives used for encoding semantic types and selectional restrictions as an unordered set. In contrast, the CLE (where such primitives are called *sorts*) organises this inventory as a *sortal hierarchy*. This hierarchy captures 28 semantic sorts. Each sort represents a class of entities. The hierarchy encodes subsumption relations between sorts. Figure 1.1 shows a part of the hierarchy.

Here, the sort `living` subsumes the two sorts `animate` and `vegetable`. `animate` can be subclassified independently in two ways: `human` vs. `animal` and `male` vs. `female`. A branching indicates incompatible subclassification, while a vertical bar indicates independent alternatives. For example, `human` is incompatible to `animal` and `vegetable`, but compatible to `male` and `female`.

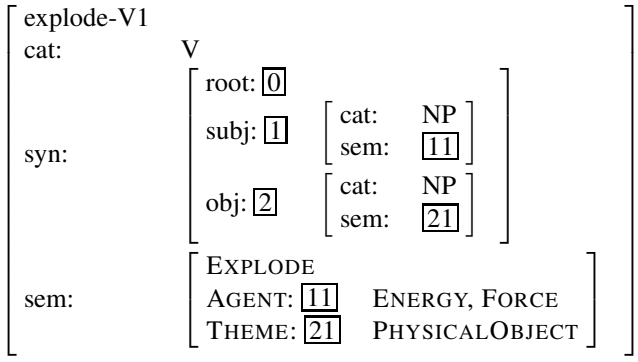
Organising semantic sorts in a subsumption hierarchy has several advantages. Subsumption relations holding between sorts can be specified once in the hierarchy instead of being redundantly repeated in various lexical entries. For example, the hierarchy states that each word classified as `human` is also classified as `animate` and `living`. Thus, it is not necessary to specify the sorts `animate` and `living` in all lexical entries bearing the sort `human`. Actually, such dependencies between sorts have to be encoded at some place, in order to cope with the fact that words satisfying certain selectional restrictions may exhibit varying levels of abstraction. Selectional restrictions which are characterised by a certain sort are also satisfied by words which belong to a subsort, i.e. are more specific. For instance, given the hierarchy in figure 1.1, the most appropriate sort for specifying the restrictions for the subject (Agent) of "creep" is `animate`. This restriction is satisfied by all nouns

of the sort human (e.g. "baby") or animal (e.g."crocodile"). Furthermore, a selectional constraint is also satisfied by words which belong to a sort subsuming the constraint-defining sort, i.e. are more general. For example, nouns of the sort living (e.g. "life form" or "being") are possible subjects of "creep", unless they belong to a subsort incompatible with animate (such as vegetable). In such cases, selectional restrictions provide additional information about the semantically underspecified complement. If someone talks about "creeping beings", one can conclude that these beings are animate. Section 2.2.2 addresses this "informational" view of selectional restrictions. These considerations show that it is important to encode subsumptional dependencies of sorts. A hierarchy is an appropriate way of modelling such dependencies.

The discussion in the previous paragraph is closely related to another crucial issue: What is the appropriate abstraction level for specifying selectional restrictions? In the example above, I stated that animate is the appropriate sort to capture the restrictions imposed by "creep". On the one hand, the more abstract sort living would be an over-generalisation, since this sort also includes plants. On the other hand, each of the two more specific sorts human and animal would be an under-generalisation. A disjunction of these sorts would adequately characterise the restriction, but this would miss a possible generalisation: one would use two sorts to encode something that one sort could express in a more compact way. Note that selecting the appropriate generalisation level crucially depends on the structure and the coverage of the employed sort hierarchy. If the sort animate were missing and living had the three immediate descendants vegetable, animal, and human instead, then the restrictions of "creep" would be best modelled by a disjunction of the latter two sorts. Moreover, one could object that the sorts offered by this hierarchy are too abstract to allow an appropriate encoding of these restrictions, because there are some kinds of animals that do not creep (e.g. fish). To capture such differentiations, a much more fine-grained semantic hierarchy would be necessary. Section 2.2.1 will discuss the linguistic finding that basically any semantic information can play a role for selectional restrictions. This finding supports the necessity of employing a comprehensive, very fine-grained hierarchy. Lexical-semantic resources like WordNet (cf. section 1.2) provide such hierarchies. However, the more abstraction levels are captured by the employed hierarchy, the more challenging is the problem of finding the appropriate generalisation level for defining selectional restrictions of a given verb. This problem will be a core issue in this thesis.

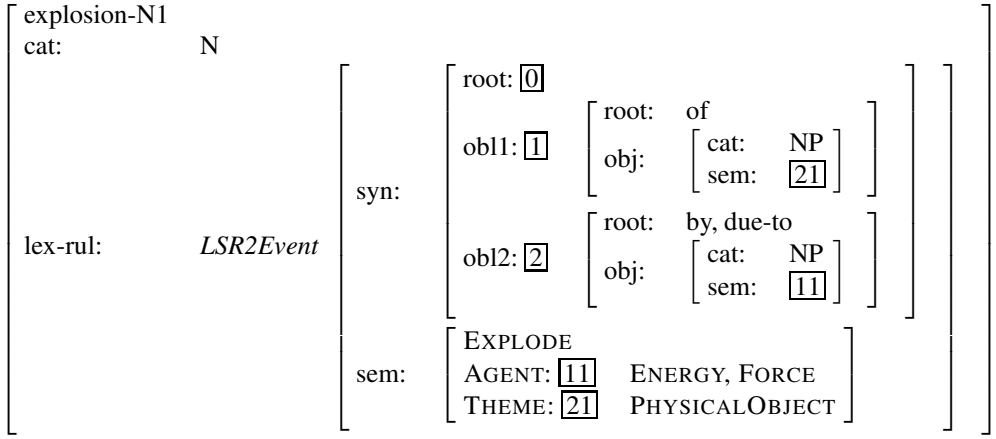
So far, when talking about selectional restrictions that predicates impose on their arguments, I have not made a clear distinction between *syntactic arguments* and their underlying *semantic arguments*. However, such a distinction is necessary, because in general, the same semantic argument can be syntactically realised in different ways. This phenomenon has been widely investigated in linguistics. In chapters 2 and 6, I will discuss these investigations in detail. In many NLP systems, lexical entries for words expressing predicates reflect this distinction by separately specifying syntactic and semantic arguments and providing links between them. For example, this strategy is pursued in the lexicon of the Mikrokosmos MT environment (cf. (Viegas 1999)). Examples (1.10) and (1.11) display two entries from this lexicon.

(1.10)



(1.10) shows the entry for "explode". This verb has two syntactic complements, specified under "syn": a subject and an object, both NPs. The two corresponding semantic arguments are specified under "sem", labelled as AGENT and THEME, respectively. These labels refer to certain *thematic roles* (or *semantic roles*). Intuitively, it is a plausible approach to identify semantic arguments by the roles they play in the event denoted by the predicate. However, a detailed determination of the nature (what is a role?) and the inventory (which roles do exist?) of thematic roles is all but trivial. Such definitions do not only vary in different NLP systems and language resources. Thematic roles have been subject to intensive linguistic research, yielding a number of competing theories. Chapter 2 will summarise the most prominent accounts. These theories also aim at finding regularities for linking syntactic arguments to their underlying semantic roles. In the lexical entry displayed here, the information which syntactic complement corresponds to which semantic role is stated by explicit links (indicated by numbers in boxes). In this case, the Agent corresponds to the subject and the Theme to the object.

(1.11)



Example (1.11) shows the lexical entry for "explosion". This entry has been generated by applying a lexical rule that accounts for nominalisations denoting events (*LSR2Event*) to (1.10). As "explosion" denotes the same kind of event as "explode", the semantic structure in the two entries are identical. However, the noun realises the semantic roles by different syntactic constituents: The Agent is expressed by a PP headed by "by" or "due to", the Theme by a PP headed by "of".

As selectional restrictions are semantic constraints, it is appropriate to attach their specification to the semantic roles rather than to the syntactic complements (cf. section 2.2.2). This is done in the entries above: the Agent is restricted to ENERGY, FORCE, the Theme to PHYSICALOBJECT. These specifications coincide in the two entries; differing syntactic realisations do not alter selectional restrictions.

In this context, a general terminological remark is necessary: In this thesis, I will use the terms *syntactic argument* and *syntactic complement* in a very broad sense. This means that I will not make a distinction between syntactic constituents which are obligatory to satisfy the valency of a verb (usually called *complements*) on the one hand and constituents which are optional modifiers of the verb (usually called *adjuncts*) on the other hand. For my work, this distinction is not crucial. Actually, thematic roles may be related to complements (in the narrow sense) as well as adjuncts; likewise, verbs may impose selectional restrictions on both kinds of verbal modifiers. For example, locative roles like Location, Source, or Goal correspond to adjuncts. The fillers of these roles are semantically restricted to denote some location.

1.2 WordNet and EuroWordNet

In general, NLP systems built till the early nineties were characterised by a very limited coverage of phenomena and regularities in natural language. In particular, their lexicons typically comprised no more than a few hundred or thousand entries. Whereas the merits of such systems consist in the investigation of useful representations and mechanisms for language processing, they are too small to be suitable for "real-world" applications, unless they are tuned to and used within a strongly delimited domain. A tool capable of processing free (i.e. unrestricted) text requires, among others, a large, broad-coverage lexicon. However, manually building such a lexicon from scratch is very labour-intensive and time-consuming. For this reason, two major developments have emerged in the last 10–15 years to overcome this lexicographic bottleneck. The first one is the development of techniques to automatically acquire lexical information from large corpora (cf. (Boguraev & Pustejovsky 1996), (Lemnitzer & Wagner 2004)). The other one is the creation of large-scale general-purpose electronic lexical resources which can be reused for arbitrary language processing systems. One of the most widespread resources of this kind is WordNet (cf. (Miller et al. 1990), (Fellbaum 1998)), which comprises lexical-semantic information for English. WordNet-like resources (usually called *wordnets*, without capital letters) for various other languages have been or are being built. EuroWordNet (cf. (Vossen 1999), (Wagner & Kunze 1999)) is a multilingual database that integrates some of these language-specific wordnets. In my work, I employ both WordNet and EuroWordNet. In the following, I introduce those aspects of the two resources that are of relevance for this thesis; an exhaustive introduction can be found in the literature just mentioned.

1.2.1 WordNet

WordNet is a lexical-semantic net capturing nouns, verbs, adjectives, and adverbs. It was developed at Princeton University and is continually being extended and improved. Basically, this resource comprises a large-scale semantic network, which defines the meaning of words via semantic relations between words and word concepts. The most elementary relation in WordNet is synonymy. Synonymous words are gathered in so-called *synsets* (synonym sets). Examples of such synsets are given in (1.12)–(1.16):

(1.12) <person#individual#someone#mortal#human#soul>

(1.13) <soul#psyche>

(1.14) <mortal#deadly>

(1.15) <drink#imbibe#take_in_liquids>

(1.16) <consume#have#ingest#take>

Each synset belongs to a certain part-of-speech: (1.12) and (1.13) are noun synsets, (1.14) is an adjective synset, and (1.15) and (1.16) are verb synsets. The ambiguity of a word is modelled by its membership in several synsets. For instance, "soul" is member of (1.12) and (1.13), "mortal" occurs in (1.12) and (1.14). Each synset in which a word occurs represents a particular sense of that word. More exactly, a synset represents an equivalence class of senses of different words. (1.12) represents a certain sense of "soul", a certain sense of "mortal", as well as a certain sense of each other word in the synset. All these senses are equivalent, i.e. they refer to the same sort of entities. To most synsets, a definitional gloss is attached which explains the meaning represented by the synset and whose primary purpose is to provide orientation to the human user. For (1.12), this gloss is "a human being". The gloss for (1.13) (representing another sense of "soul") is "the immaterial part of a person"; the one for (1.14) (representing another sense of "mortal") is "causing or capable of causing death". Representing word senses is one of the fundamental functions of synsets.

Synsets form the nodes of the semantic network encoded in WordNet. These nodes are connected by several types of semantic relations like the following:

- hyponymy/hyperonymy (subordinate/superordinate; <person>[3] is a hyponym of <life_form>)
- antonymy (opposite; <lightness> and <darkness> are antonyms[4])
- meronymy/holonymy (part/whole; <hand> is a meronym of <arm>)
- entailment (implication; <succeed> entails <try>)
- cause (<kill> causes <die>)

This list is not exhaustive.

Such relations form the edges of the semantic network. Essentially, there is a separate network for each part-of-speech, i.e. most relations hold between synsets of the same part-of-speech. However, certain relation types establish connections between these networks (e.g. expressing derivation). In the different part-of-speech networks, different relation types are prevalent. For the nouns, the relation type which is encoded in extenso is hyponymy/hyperonymy. These relations constitute a subsumption hierarchy over the noun synsets, which is comparable to the CLE hierarchy discussed in section 1.1.2 (cf. figure 1.1 on page 4).[5] Figure 1.2 displays a part of that hierarchy. This figure shows that <animal> has (among others) the synsets <insectivore>, <pet>, and <invertebrate> as hyponyms, <life_form> as hyperonym, and (among others) <human> and <plant> as co-hyponyms (siblings). The hyponymy/hyperonymy relation can be interpreted transitively; in this sense, <insectivore>,

[3] If the context provides sufficient information about a synset, I will not mention all synset members.

[4] I will not refer to the difference between lexical and conceptual relations here.

[5] Note that the WordNet hierarchy does not distinguish between compatible and incompatible co-hyponyms as does the CLE hierarchy. The EuroWordNet database schema provides the possibility to encode this distinction. However, this is not realised for all wordnets in EuroWordNet.

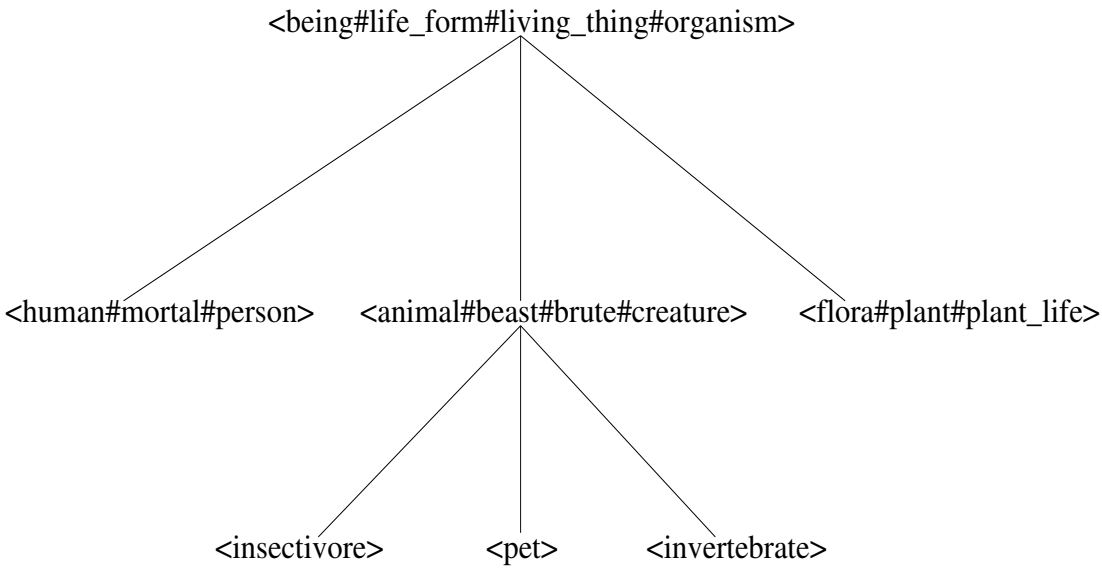


Figure 1.2: Part of the WordNet noun hierarchy

<pet>, and <invertebrate> are (indirect) hyponyms of <life_form>. This leads us to a second fundamental function of synsets: they do not only represent word senses, but also semantic concepts which are abstractions over word senses. For instance, <life_form> represents a certain sense of "being", "life form", "living thing", and "organism", but also an abstract concept subsuming senses of a number of other words, such as "person", "animal", "insectivore", "pet", "invertebrate", or "plant". It will turn out that the acquisition approach I pursue in this thesis requires that these two fundamental functions are representationally separated. From now on, I will use the term *concept* to refer to a synset. This conforms to the terminology used in EuroWordNet.

One of the crucial properties that renders WordNet suitable for real-world applications is its size. Version 1.5, which I use for my work[6] comprises more than 60 000 noun concepts, 11 000 verb concepts, 16 000 adjective concepts, and 3 000 adverb concepts.

1.2.2 EuroWordNet

Since the time when WordNet has become available, a large community of researchers and developers has been employing this resource for various NLP applications, including, but not limited to word sense disambiguation (cf. (Stetina, Kurohashi & Nagao 1998)), semantic annotation of corpora (cf. (Miller, Leacock, Tengi & Bunker 1993)), information retrieval (cf. (Gonzalo, Verdejo, Chugur & Cigarran 1998)), and text categorisation (cf. (Buenaga Rodriguez, Gomez-Hidalgo & Diaz-Agudo 1997)); cf. (Kunze & Wagner 2001) for a survey. In parallel, numerous initiatives have emerged to create a wordnet in other languages than English; currently, wordnets in more than 30 languages exist or are under development. These initiatives are brought together in the Global WordNet Association (GWA)[7]. This organisation aims at providing a platform for establishing cooperations concerning wordnet creation and dissemination, developing wordnet-related standards, and promoting the integration of language-specific wordnets in multilingual resources. For the latter goal, data and strategies will be employed that have been developed for the first existing multilingual wordnet database, EuroWordNet.

EuroWordNet (EWN) was created from 1996 to 1999 and comprises wordnets for eight European languages: Czech, Dutch, English, Estonian, French, German, Italian, and Spanish. Currently, EWN is restricted to nouns and verbs. The language-specific wordnets are integrated in a common database. This database has a modular structure. Each wordnet with its concepts and relations is stored separately. The inventory of possible relations (e.g. HAS_HYPONYM, HAS_HYPERONYM, ANTONYM, etc.) is defined globally; it captures all relations which occur in any of the individual wordnets. The core of the EWN database is the Interlingual Index (ILI). The language-specific wordnets are aligned via that index. Basically, the ILI is an unstructured set of records representing concepts. For practical reasons, the noun and verb concepts of WordNet (version 1.5) have been adopted for this set. A concept in a language-specific wordnet is linked to a corresponding ILI record via a so-called equivalence relation. In this way, the ILI forms a "bridge" that allows mappings between different wordnets.

Figure 1.3 illustrates the general structure of EuroWordNet. The language-specific wordnets are grouped around the ILI. In the example, concepts from different wordnets are linked to the ILI record <drive>. Thus, the interconnection to the ILI allows indirect mappings between these language-specific concepts. In addition to the ILI, two language-independent ontologies are encoded in EWN:

[6] The latest WordNet version is 2.0. I employ version 1.5 for the sake of compatibility, because it is part of EuroWordNet.

[7] http://www.globalwordnet.org

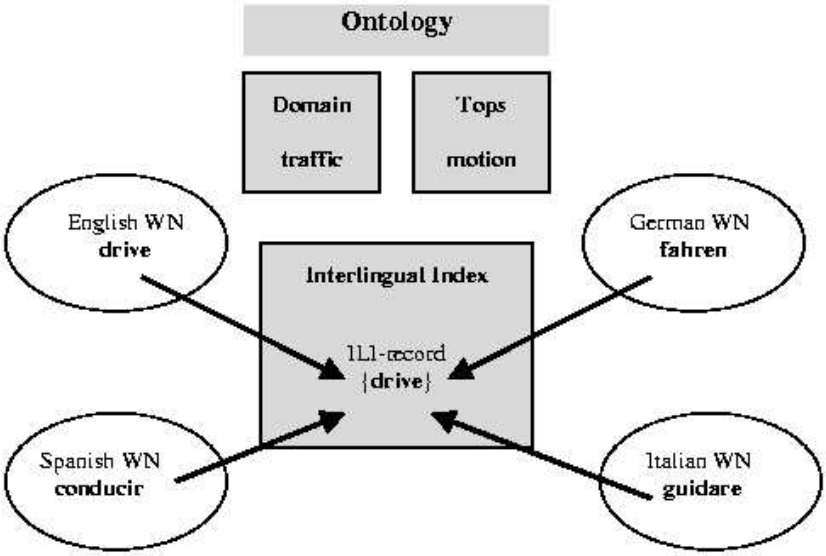


Figure 1.3: General architecture of EuroWordNet

a top ontology, consisting of a small set of abstract semantic features such as ARTEFACT, ANIMAL, DYNAMIC, or EXISTENCE, and a (currently incomplete) domain ontology encoding specific semantic fields (domains) such as computer terminology. These ontologies are used for classifying ILI concepts. They do not play a role in this thesis.

The modular architecture of EuroWordNet allows the independent creation and integration of the different language-specific wordnets. Basically, two strategies have been applied for that task. Some wordnets have been built up from scratch, and equivalence relations to the ILI have been encoded afterwards (*merge approach*). This approach has been pursued, for instance, for the German wordnet (GermaNet), cf. (Kunze & Wagner 2001), (Wagner & Kunze 1999). Other wordnets (e.g. the French wordnet) have been built by creating a translation of Princeton WordNet, i.e. by semi-automatically translating WordNet concepts into equivalent concepts in the respective language and adopting the relations from WordNet which connect these concepts (*expand approach*). Since the ILI itself consists of WordNet concepts, this strategy yields links to the ILI records as a by-product: each synset is linked to its WordNet origin. Both approaches have advantages and drawbacks. Building a wordnet from scratch allows the adequate modelling of lexical-semantic patterns specific to the respective language, while the adoption of the WordNet structure could miss such patterns and yields biases peculiar to English (and to Princeton WordNet). On the other hand, creating a translation of WordNet is much less labour-intensive, since it can be done semi-automatically and there is no additional effort to encode equivalence relations to the ILI.

In order to ensure a certain degree of homogeneity of concepts and semantic fields covered in the different wordnets, the EuroWordNet partners agreed on a set of *common base concepts* in the ILI, which are characterised by a high level of abstraction (without being at the top level of the hierarchy), a high number of relations, and high frequency in corpora. Each wordnet has to encode counterparts of all these base concepts, and these counterparts have to be linked to the corresponding ILI records.

In case of the merge approach (a wordnet is created independently from WordNet), it often happens that for some language-specific concept there is no exact counterpart in the ILI. This could be due to missing lexicalisation in English or due to a gap in WordNet. To account for such cases, several types of equivalence relations are defined that allow flexible linking. (1.17)–(1.22) provide some example equivalence links from the German wordnet to the ILI:

(1.17) <Weihnachten> EQ_SYNONYM <Christmas>

(1.18) <Allerheiligen> EQ_SYNONYM <Allhallows>

(1.19) <Feiertag> EQ_NEAR_SYNONYM <day>

(1.20) <Feiertag> EQ_NEAR_SYNONYM <holiday>

(1.21) <Buß-_und_Bettag> EQ_HAS_HYPERONYM <Christian_holy_day>

(1.22) <Buß-_und_Bettag> EQ_HAS_MERONYM <penitence>

Full correspondence is indicated by the EQ_SYNONYM link, as in (1.17) and (1.18). This is the usual equivalence relation which should be used whenever possible. Sometimes a language-specific concept is (only) roughly synonymous to an ILI record. In this case, the EQ_NEAR_SYNONYM link is used. In particular, this relation is employed if two or more ILI concepts correspond to one wordnet concept or vice-versa. Such a case is shown in (1.19) and (1.20). The meaning of the German concept <Feiertag> is captured to some extent by the WordNet concept <day> (with the gloss "a day assigned to a particular purpose or observance") subsuming concepts like <Valentine's_Day>, <First_of_May>, or <Mother's_Day>, and to some extent by <holiday>, which does not completely match <Feiertag> because it does not subsume religious holidays in WordNet. (1.21) and (1.22) show the links for <Buß-_und_Bettag>, a protestant holiday in Germany devoted to penitence. For this holiday, there is no corresponding ILI record. Therefore, the meaning is modelled by an EQ_HAS_HYPERONYM link to <Christian_holy_day> and an EQ_HAS_MERONYM link to <penitence>.

The size of the individual wordnets in EWN differ significantly; they range from 7 500 concepts for Estonian to 30 000 concepts for Dutch, Italian, and Spanish. For English, WordNet 1.5 is integrated, supplemented by an English wordnet encoding additional (mostly derivational) relations.

1.2.3 Selectional Restrictions in EuroWordNet

Although, as we have seen in section 1.1, selectional restrictions are very useful for language processing tasks, WordNet does not encode such information at all. However, some of the wordnets in EWN comprise relations which capture selectional restrictions. These relations encode thematic roles. They provide links between verb concepts and noun concepts. Essentially, they specify what kind of noun concepts are involved, i.e. play a role, in events denoted by verb concepts. The relations are bi-directional; the direction pointing from verb to noun is labelled INVOLVED, the reverse direction from noun to verb is labelled ROLE.[8] Several types of such thematic role relations are defined, e.g. AGENT, PATIENT, or INSTRUMENT. (1.23)–(1.27) display some examples of thematic role relations from the Italian wordnet. The ILI records which correspond to the Italian concepts are provided as well.

[8]For each relation, the encoder specifies one direction; the reverse direction is added automatically.

(1.23) <abbeverarsi> INVOLVED_AGENT <animale>
<drink> *<animal>*

(1.24) <camminare> INVOLVED_AGENT <pedone>
<walk> *<pedestrian>*

(1.25) <lessare> INVOLVED_PATIENT <cibo>
<boil> *<food>*

(1.26) <sbaraccare> INVOLVED_PATIENT <cosa>
<remove> *<object#inanimate_object>*

(1.27) <lavare> INVOLVED_INSTRUMENT <acqua>
<wash> *<water>*

Thematic role relations encode selectional restrictions in an analogous way as shown in the lexical entry (1.10) on page 6 from the Mikrokosmos lexicon (cf. (Viegas 1999)), i.e. as relations between semantic concepts which are characterised by similar labels (e.g. AGENT). However, there are two differences between the representation in that lexicon and in EuroWordNet: Firstly, thematic role relations are not encoded exhaustively in the EWN wordnets. Those wordnets which contain such relations encode them only for a minority of verb concepts. Furthermore, these wordnets in general do not provide role relations for *all* arguments of a verb concept, but, if at all, only for *some* arguments (in many cases for just one). Secondly, EWN does not provide information about the possible syntactic realisations of semantic arguments, as does the Mikrokosmos lexicon. The EuroWordNet database scheme allows to specify syntactic subcategorisation frames of verbs. Such specifications are realised in the different wordnets to different degrees. For example, WordNet attaches coarse characterisations of possible subcategorisation frames to verb concepts, e.g. "Something __s" for an intransitive and "Something __s something" for a simple transitive frame. However, none of the wordnets in EWN captures connections between syntactic and semantic arguments.

1.3 The Task of this Thesis

At the beginning of section 1.2, I stated the two main strategies pursued in the last decade to overcome the lexical acquisition bottleneck: learning lexical information from corpora and building reusable lexical resources. These strategies can be combined. For example, (Biemann, Bordag & Quasthoff 2004) present a method to extend wordnets by automatically acquiring paradigmatic relations (synonyms, hyperonyms, co-hyponyms) by means of statistical analysis of co-occurrence data in large corpora. The words and relations acquired in this way are candidates for the extension of wordnets. They have to be inspected manually before being added to the respective wordnet.

This thesis also deals with the (semi-)automatic extension of lexical semantic resources. In particular, it focuses on wordnets. The task of this work is to develop an approach for learning thematic role relations (as described in section 1.2.3) from corpora. As shown in this chapter, thematic role relations are intrinsically related to selectional restrictions. Consequently, the core of a strategy for learning role relations is a strategy for learning selectional restrictions. To fulfil the latter task, a number of methods have been proposed. These methods take syntactically analysed corpora as input. More specifically, they act on verbs and their syntactic arguments, which are identified by such corpora. To acquire the selectional preferences of a certain verb, the complements of that verb are extracted from the corpus,

and statistical methods and heuristics are applied to generalise over these complements; this generalisation is expressed as a set of noun concepts in WordNet. I will discuss such approaches in chapter 3. They constitute the starting point of my research. However, they are not immediately applicable for learning role relations but require enhancements and extensions for that task. In particular, two aspects have to be taken into account:

- The WordNet concepts acquired to represent selectional preferences of a verb have to be located at an appropriate level of generalisation. I have discussed that problem in section 1.1.2. I will argue that the approaches presented so far do not behave satisfyingly in this respect.
- As the existing methods generalise over syntactic complements of verbs, they acquire selectional preferences for *syntactic* rather than *semantic arguments*. To learn selectional restrictions for semantic roles, the syntactic arguments provided by the parsed corpus have to be mapped to their underlying roles. Then, the statistical methods and heuristics can be applied, say, to all Agents of the examined verb rather than to all its subjects. Therefore, the task of this thesis requires to develop a method for linking syntactic to semantic arguments.

It is an important goal that the learning approach developed in this thesis is applicable in a wide range of environments. Therefore, it should meet the following criteria:

- *Language independence:* The approach should be as language-independent as possible, so that it can be applied to extend wordnets of any language. In particular, this implies that only two kinds of resources may be employed: the wordnet to be extended and a syntactically analysed corpus in the respective language. The approach should not depend on further resources, e.g. lexicons from which thematic relations could be extracted, because such resources are not available for many languages. I will develop and test my approach with English data, i.e. with Princeton WordNet and a huge parsed English corpus. However, it will be applicable for other languages as well.
- *Theory neutrality:* As already mentioned, there is no consensus in linguistics about the nature and the inventory of thematic roles. In chapter 2, I will discuss a number of linguistic theories of thematic roles. The approach I develop should be open to different theories, i.e. it should not depend on a particular theory. Nevertheless, it should take into account general (undisputed) linguistic insights concerning thematic roles and selectional restrictions as far as possible. In this thesis, I will adopt the role inventory and definitions assumed in EuroWordNet. However, this decision is guided by purely practical reasons: To evaluate my approach, I will extract a gold standard from EuroWordNet and compare the acquired role relations with this gold standard. To make this comparison possible, the acquired types of relations have to correspond to those relation types offered by EWN. However, the overall approach does not presuppose the adoption of these particular role types.

Regarding the extension of wordnets, I generally assume the scenario mentioned at the beginning of this section: the learning approach I propose supplies candidate relations, which have to be manually examined before being added to the respective wordnet. In this sense, I propose a semi-automatic approach for enriching a wordnet with thematic role relations. However, one could also imagine scenarios where the new role relations are not intended to be persistently stored in the general, reusable version of a wordnet, but to be acquired by a particular user to support a specific NLP application. In this case, manual post-editing might be unnecessary for the given task.

Although this work focuses on the extension of wordnets, the learning approach to be developed can in principle be used for enriching other lexical resources—e.g. lexical semantic nets like FrameNet (cf. (Fillmore, Wooters & Baker 2001)) or lexicons employed in parsing systems—with information about selectional preferences and/or thematic role relations as well. In general, two possibilities are conceivable: If the resource to be extended provides a semantic hierarchy of noun concepts, then the method proposed in this thesis can be applied straightforwardly using this hierarchy instead of WordNet. If, however, the respective resource does not include such a hierarchy, then selectional preferences or thematic role relations can be acquired using the noun hierarchy of WordNet (or another wordnet, respectively) and the resulting WordNet concepts can be mapped to equivalent representations in the resource (i.e. words, semantic primitives, etc.).

In conclusion, I would like to summarise the essential issues which are addressed in this thesis:

- The goal of this work is to develop an approach for learning thematic role relations. This approach takes WordNet and a parsed corpus as input and acquires relations as shown in (1.23)–(1.27).
- A core component of this approach is a method of acquiring selectional preferences. In particular, these selectional preferences have to model an appropriate level of generalisation.
- Another core component is a strategy for mapping syntactic arguments, which are encoded in the training corpus, to their underlying semantic arguments, which represent thematic roles.
- The approach should be independent from a particular language or a particular linguistic theory. Nonetheless, it should reflect general linguistic insights concerning thematic roles and selectional restrictions.
- The statistical methods employed and developed for the acquisition approach should be theoretically sound.
- To evaluate the approach, I will use a gold standard, which is extracted from the EuroWordNet database. In addition to measuring the performance of the learning approach, the evaluation experiments will also assess the general feasibility of employing a gold standard for evaluating the task of learning thematic role relations, and reveal strengths and weaknesses of the particular gold standard retrieved from EWN.

1.4 Outline of this Thesis

The rest of this thesis is structured as follows: Chapter 2 provides a summary of different linguistic theories concerning thematic roles and selectional restrictions. This chapter also states which aspects of these theories are relevant for the task of this thesis and the strategy I will propose for its solution. In this context, I will give an introduction of the role inventory used in EuroWordNet and discuss the notion of selectional preference. Chapter 3 introduces several approaches for acquiring selectional preferences proposed in the literature. In particular, I will discuss their suitability concerning the task of this thesis. I will select one of these approaches as a starting point for developing my own method. Chapter 4 reports preliminary experiments testing this approach. These experiments show that this strategy exhibits an inherent weakness regarding the selection of the appropriate generalisation level for encoding selectional preferences. I will explain the reason underlying this weakness and propose

a modification of the approach that overcomes this drawback. Chapter 5 addresses several practical issues concerning the method of learning selectional preferences. These issues include the employed training data, lexical disambiguation of the words in these data, the mapping of word forms to concepts in WordNet, and the compilation of a gold standard for evaluation. First evaluation experiments round off this chapter. In chapter 6, I develop a strategy for mapping syntactic arguments extracted from the training corpus to their underlying thematic role types. This strategy makes it possible to apply the algorithm for learning selectional preferences to thematic roles. Chapter 7 reports experiments performing a detailed evaluation of the overall approach developed in this thesis and particular aspects of it. Chapter 8 provides concluding remarks.

Chapter 2

Linguistic Foundations

Whereas the previous chapter concentrated on *computational linguistic* aspects concerning the task of this thesis, this chapter addresses the relation between this task and *theoretical linguistic* research on thematic roles and selectional restrictions. Numerous theoretical accounts have been proposed in linguistics to appropriately model these phenomena. Neither the representation of thematic role relations in EuroWordNet nor the approach for learning such relations developed in this work adhere strictly to any of these theories. However, both have to a large extent been inspired by fundamental linguistic insights and concepts formulated within this area. On the other hand, it turns out that certain features of my learning approach, though motivated by independent reasons, can be interpreted in terms of linguistic theories and considerations. In my view, it is important to clarify in which respects the approach I propose conforms to or deviates from established linguistic theories. Consequently, throughout the presentation of my work in the subsequent chapters, I will explicitly refer to linguistic notions which correspond with particular aspects of my approach. However, these notions can hardly be understood in isolation, i.e. separated from their theoretical context. Therefore, this chapter provides a self-contained summary of those theories I will make reference to. Section 2.1 introduces linguistic theories on thematic roles. Section 2.2 turns to linguistic views of selectional restrictions and their relatedness to thematic roles. To give a preliminary idea of the overall relevance of these theories for the task of this thesis, these sections also indicate connections to my approach as well as to NLP implementations discussed in chapter 1. Section 2.3 addresses the significance of linguistic insights for this thesis in a more general manner. In particular, section 2.3.1 relates the ideas presented in section 2.1 and 2.2 to the approach of encoding thematic roles in EWN. Section 2.3.2 provides a linguistic motivation of the notion of selectional preferences, which I adopt for my work.

2.1 Thematic Roles

Thematic roles, also called *thematic relations*[1] or *θ-roles*, are characterisations of certain semantic relationships which hold between a verb and its complements (and adjuncts). As we have seen in chapter 1 (example (1.10) on page 6; examples (1.23)–(1.27) on page 13), these relationships are usually indicated by labels like *Agent*, *Theme*, or *Goal*, which are assigned to the arguments of a verb.

[1] To avoid confusion between the notion *thematic relation* as a linguistic phenomenon and the technical term *relation* as a structural part of wordnets, I will use *thematic role* to refer to the linguistic concept and *thematic role relation* to refer to a wordnet relation which encodes a thematic role.

For example, in the sentence

(2.1) Peter sent the products from London to New York.

"Peter" is the *Agent*, "the products" the *Theme* or *Patient*, "from London" the *Source*, and "to New York" the *Goal* of the sending event denoted by the sentence.

Up to now, linguistic research has not come to a consensus about the exact inventory of thematic roles, nor about their nature or their status in linguistic theory. Roles like those just enumerated were first mentioned by Jeffrey Gruber, who provided a detailed study about syntactic and semantic observations connected with them (cf. (Gruber 1965)). The most prominent theoretical accounts in this domain have been developed by Charles Fillmore, Ray Jackendoff (who, to a large extent, takes up Gruber's observations and conclusions), and David Dowty. In the following sections, I will sketch the main results of their work.

2.1.1 Fillmore's Case Grammar

Fillmore does not use the term *thematic roles* directly. However, in (Fillmore 1968) he introduces the very similar concept of *deep cases*, which he defines as "semantically relevant syntactic relationships involving nouns and the structures that contain them", i.e. superordinate predicates (Fillmore 1968, p. 5).[2] Deep cases make explicit the semantic functions of the complements of a predicate.

Fillmore provides a (non-exhaustive) list of such deep cases (Fillmore 1968, p. 24–25):

Agentive (A) the case of the typically animate perceived instigator of the action identified by the verb

Instrumental (I) the case of the inanimate force or object causally involved in the action or state identified by the verb

Dative (D) the case of the animate being affected by the state or action identified by the verb

Factitive (F) the case of the object or being resulting from the action or state identified by the verb, or understood as a part of the meaning of the verb

Locative (L) the case which identifies the location or spatial orientation of the state or action identified by the verb

Objective (O) the semantically most neutral case, the case of anything representable by a noun whose role in the action or state identified by the verb is identified by the semantic interpretation of the verb itself; conceivably the concept should be limited to things which are affected by the action or state identified by the verb.

Note that these definitions include or imply semantic restrictions for individual cases: A and D are (typically) animate, I is inanimate, L denotes some location. I will use such restrictions as semantic

[2] In (Fillmore 1977), he also uses the term *semantic roles*.

filters which form a component of my strategy of mapping syntactic arguments to their underlying thematic roles (cf. section 6.5). Such filters help to distinguish between different role types. For example, according to the definitions above, a subject that is inanimate cannot be Agentive.

A deep case can be syntactically realised in various ways. Recall that the lexical entries (1.10) and (1.11) in section 1.1.2 illustrate this fact. These entries pertain to a verb and its nominal derivation, respectively. However, a verb itself may allow several realisations of its semantic arguments. For example, look at (2.2)–(2.4):

(2.2) John broke the window.

(2.3) A hammer broke the window.

(2.4) John broke the window with a hammer.

In (2.2) and (2.4), the subject is the Agentive, whereas in (2.3), it is the Instrumental. In (2.4), in contrast, the Instrumental is realised by a *with*-PP. As sentence (2.5) shows, the subjects of (2.2) and (2.3) cannot be conjoined:

(2.5) *John and a hammer broke the window.

This shows that they must differ with respect to their semantic relation to the verb. Following (Fillmore 1977) such examples are useful for discriminating deep cases: If two disparate classes of nominals occur in the same syntactic relation to a certain verb, as in (2.2) and (2.3), this may indicate that this verb co-occurs with two different deep cases which can be realised by this syntactic relation. Further strong evidence for that is provided by sentences in which the two nominal classes both co-occur with the verb in question, but in different syntactic positions, as in (2.4).

Fillmore mentions dependencies between the cases which occur in a sentence. For example, if an active transitive sentence contains an Instrument in a *with*-PP (as in (2.4), then its subject must be a personal Agent. He explains the apparent counter-example

(2.6) The car broke the window with its fender.

by pointing out that it corresponds to

(2.7) The car's fender broke the window.

since the determiner of "fender" in (2.6) has to be a possessive pronoun referring to "car":

(2.8) *The car broke the window with a fender.

For practical reasons, the approach I develop in this thesis (as well as comparable approaches discussed in chapter 3) does not take into account semantic dependencies between different arguments (cf. section 3.2).

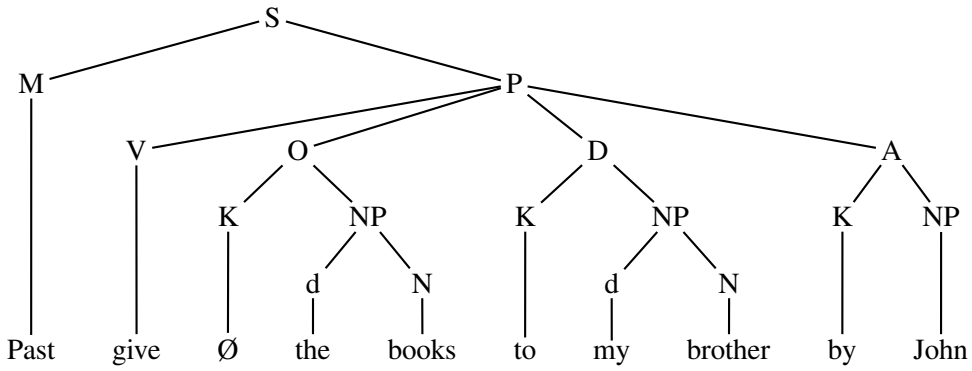
The set of case categories (the *case frame*) which co-occur with a particular verb is specified in its lexical entry. For example, "die" co-occurs with a Dative, whereas "kill" co-occurs with a Dative, an optional Instrument, and an Agent:

(2.9) *die* (+[_____ D])

(2.10) *kill* (+[_____ D (I) A])[3]

Formalising his theory within the transformational grammar paradigm, Fillmore claims that the deep structure of a sentence contains a P (Proposition) constituent which comprises the verb and its complement constituents, each of which is labelled with the appropriate case category. (2.11) is an example of a deep structure (M is the modality constituent, which we neglect here; K is the case marker, which is realised as a preposition in English[4]).

(2.11)



By certain transformations (among others, the movement of one case constituent to the subject position and the deletion of the preposition of that constituent), the following sentences can be derived from (2.11):

(2.12) John gave the books to my brother.

(2.13) John gave my brother the books.

(2.14) The books were given to my brother by John.

(2.15) My brother was given the books by John.

For "give" in active voice, A has to be realised as subject (if an Agent occurs in the sentence), while either O or D may move to the direct object position (where the preposition is deleted as well). O or D can only move to the subject position if the verb is passivised. Thus, in the "normal" (unmarked) configuration, the Agent is the subject. Fillmore states that there is a general, verb-independent principle for the "unmarked" subject choice, the *subject selection principle*:

[3]Formally, the notation here represents a contextual feature. This feature specifies the context (i.e. the syntactic configuration) in which the respective lexical item may be inserted.

[4]Fillmore points out that prepositions in English or postpositions in Japanese can be viewed as analogous to morphological case affixes in languages like Latin. Therefore, he proposes a broader definition of the notion *case*, to capture also the marking of verbal complements by prepositions or postpositions.

> If there is an A, it becomes the subject; otherwise, if there is an I, it becomes the subject; otherwise, the subject is the O. (Fillmore 1968, p. 33)

A transformational grammar comprises transformations which generate the surface structure from the deep structure. Thus, in Fillmore's account, transformations map deep cases to surface syntactic arguments. The strategy of linking syntactic complements to thematic roles, which is a crucial module of my acquisition approach, establishes a mapping in the reverse direction. Nonetheless, findings of the kind just mentioned, i.e. statements specifying relationships between surface syntactic arguments and deep cases / thematic roles, form the base of stipulating heuristics which are one component of my linking strategy (cf. section 6.3 and 6.4).

In (Fillmore 1977), he provides a more elaborated account of assigning semantic arguments of a verb to the subject and the object position (*argument selection*). He argues that any linguistic expression is connected to a specific situation, a *scene* (background) and, moreover, brings certain aspects of this situation into *perspective* (foreground). For example, both verbs "buy" and "sell" refer to the commercial situation of exchanging goods for money, but "buy" refers to it from the perspective of the buyer, "sell" from the perspective of the seller. Fillmore claims that just those elements of the situation which are to be taken into perspective are realised as subject or direct object. Hence, while "buy" puts the buyer and the goods into the foreground, "sell" emphasises the seller and the goods, and "spend" the buyer and the money.

Fillmore observes that some elements of a scene are more salient than others and thus preferably taken into perspective. There are certain semantic conditions which determine this preference, and hence the selection for the subject or the object position. This set of conditions, which Fillmore calls *saliency hierarchy*, includes humanness, undergoing a change, definiteness, and totality. For example,

(2.16) I hit Harry with the stick.

("Harry" is object) is more natural than

(2.17) I hit the stick against Harry.

("the stick" is object) because Harry is human. Likewise, in

(2.18) He knocked the door down.

("the door" is object) the door undergoes a change, but in

(2.19) He knocked on the door.

("the door" is inside a PP) it does not. Verbs of the widely examined *spray/load* class (cf. (Levin 1993)) exhibit the following alternation:

(2.20)

a. I sprayed paint onto the wall.[5]

b. I loaded hay onto the truck.

(2.21)

a. I sprayed the wall with paint.

b. I loaded the truck with hay.

The sentences in (2.21) imply completeness: the wall is completely covered with paint, and the truck is completely loaded with hay. The sentences in (2.20) do not have this implication. Thus, the saliency criterion of totality (completeness) favours putting "wall" or "truck" at the object position.

Fillmore's approach of explaining the correspondence between syntactic and semantic arguments in terms of foreground, background, and saliency is not applicable for the task of this thesis. The data which the acquisition algorithm takes as input—a parsed corpus and WordNet—do not provide that kind of situational information.

2.1.2 Jackendoff's Conceptual Semantics

As mentioned at the beginning of section 2.1, Jackendoff's research of thematic roles (cf. (Jackendoff 1987), (Jackendoff 1990)) is based on empirical observations and conclusions of (Gruber 1965). In particular, he starts from Gruber's inventory of roles. This inventory rests on the insight that "the formalism for encoding concepts of spatial location and motion, suitably abstracted, can be generalised to many other semantic fields. The standard evidence for this claim is the fact that many verbs and prepositions appear in two or more semantic fields, forming intuitively related paradigms." (Jackendoff 1990, p. 25) This fact is illustrated by the following examples:

(2.22)

a. *Spatial location and motion*

   i. The bird went from the ground to the tree.

   ii. The bird is in the tree.

   iii. Harry kept the bird in the cage.

b. *Possession*

   i. The inheritance went to Philip.

   ii. The money is Philip's.

   iii. Susan kept the money.

d. *Ascription of properties*

   i. The light went/changed from green to red.

   ii. The light is red.

[5] Fillmore uses a similar example with the verb "smear", which also belongs to the *spray/load* class.

iii. Sam kept the crowd happy.

d. *Scheduling of activities*

i. The meeting was changed from Tuesday to Monday.

ii. The meeting is on Monday.

iii. Let's keep the trip on Saturday.

The (i.) sentences express a change, the corresponding (ii.) sentences the state that results from that change, and the (iii.) sentences the causation of an enduring state.

Based on this observation, the thematic roles introduced by Gruber and adopted by Jackendoff originate, in their literal sense, from the spatial semantic field:

Theme the object in motion or being located

Source the object from which motion proceeds

Goal the object to which motion proceeds

Agent the instigator of an action or state

We saw a sentence with exactly these roles in example (2.1).

Jackendoff proposes several modifications and refinements of Gruber's role inventory. But before we can turn to this issue and to Jackendoff's analyses of thematic roles in general, I have to sketch the theoretical framework within which these analyses are formalised, namely Jackendoff's *conceptual semantics*.[6]

Jackendoff represents the meaning of a linguistic expression by a *conceptual structure*. A conceptual structure is a composition of *conceptual constituents*. Each conceptual constituent belongs to a *major ontological category* (a "conceptual part-of-speech") like Thing, Event, State, Place, Path, or Property. A conceptual constituent essentially comprises one or more atomic semantic primitives or function-argument structures. A function is a semantic primitive like GO, BE, CAUSE, TO, FROM, etc. The arguments are themselves conceptual constituents of a certain ontological category. For example, the basic conceptual structures that are expressed by "go", "be", and "keep", respectively, in (2.22) are

(2.23)

i. $\left[_{\text{Event}}\ \text{GO}\ (\left[\ \ \right],\ \left[_{\text{Path}}\ \begin{matrix}\text{FROM}\ (\left[\ \ \right])\\ \text{TO}\ (\left[\ \ \right])\end{matrix}\ \right])\right]$

ii. $\left[_{\text{State}}\ \text{BE}\ (\left[\ \ \right],\ \left[_{\text{Place}}\ \ \right])\right]$

iii. $\left[_{\text{Event}}\ \text{STAY}\ (\left[\ \ \right],\ \left[_{\text{Place}}\ \ \right])\right]$

[6] In contrast, Gruber, like Fillmore, worked within the paradigm of transformational grammar, although the syntactic structures he proposes differ significantly from Fillmore's.

For instance, (2.23 i.) represents the function GO which takes two arguments, and the second argument contains a Path comprising a FROM and a TO function.[7]

The conceptual structure for a sentence is a composition of the conceptual structures for its parts. An important requirement that conceptual semantics poses for the mapping between a syntactic structure and a conceptual structure is that for every major constituent in the syntactic structure there has to be a corresponding constituent in the conceptual structure, as illustrated by the following example:

(2.24)

a. [S [NP John] [VP ran [PP into [NP the room]]]]

b. [Event GO ([Thing JOHN], [Path TO ([Place IN ([Thing ROOM])])])]

(2.24 b.) is the conceptual structure which corresponds to the syntactic structure (2.24 a.). In particular, the verb corresponds to the function GO, the subject to the first argument of GO, the PP to the second argument of GO, and the NP inside the PP to the argument of IN. Note that the reversal of this condition is not necessarily true: there may be conceptual constituents that do not have a syntactic counterpart, like the Place constituent here.

The mapping between syntactic and conceptual structures is governed by correspondence rules. Correspondence patterns which are specific for particular words have to be encoded in the respective lexical entries like

(2.25)
$$\left[\begin{array}{l} \textit{run} \\ \text{V} \\ \underline{\qquad}\langle \text{PP}_j \rangle \\ [\text{Event GO } ([\text{Thing }]_i, [\text{Path }]_j)] \end{array}\right]$$

(2.26)
$$\left[\begin{array}{l} \textit{into} \\ \text{V} \\ \underline{\qquad}\text{NP}_j \\ [\text{Path TO } ([\text{Place IN } ([\text{Thing }]_j)])] \end{array}\right]$$

A lexical entry provides the syntactic subcategorisation frame (the angle brackets indicate optional complements) as well as the conceptual structure corresponding to the word. Indices mark correspondences between syntactic complements and conceptual substructures. (Index i refers to the subject, which is not explicitly mentioned in the subcategorisation frame.) The appropriate combination of the conceptual structures of the words in a sentence yields its complete conceptual structure.

Jackendoff formalises the nature of thematic roles in terms of conceptual semantics. He claims that thematic roles are structural relations within conceptual structures. In other words, a role label is just an abbreviatory term for a specific structural configuration in a conceptual structure. In this way, the thematic roles mentioned above are defined as:

[7] To distinguish between the different semantic fields, an additional feature is added, which Jackendoff notates as a subscript on the main function. Thus, for the semantic fields mentioned above, we have $\text{GO}_{\text{Spatial}}$, GO_{Poss}, GO_{Ident}, and GO_{Temp}, respectively.

Theme the first argument of any of the location or motion functions (GO, BE, STAY, etc.)

Source the argument of FROM

Goal the argument of TO

Agent the first argument of CAUSE

This approach of defining thematic roles in terms of semantic structure instead of representing them as unanalysed labels (as done e.g. by Gruber and Fillmore and in the GB literature) has several advantages. The semantics of a role does not have to be stipulated but is motivated independently. Furthermore, this semantics is much clearer, avoiding vague conceptions like a "default" role Objective in Fillmore's case grammar. Analogously, relationships between different roles become explicit (in terms of conceptual configurations), which is not the case if we just have an unstructured list of role types. Thus, the containment and delimitation of thematic roles is much more straightforward, and hence determining an inventory of thematic roles is facilitated.

Despite these advantages, EuroWordNet represents thematic roles by relations bearing unanalysed labels. The reason for this is that the apparatus for semantic representation in wordnets differ fundamentally from the one in Jackendoff's approach. Jackendoff represents the meaning of words and sentences by the composition of semantic primitives, whereas wordnets encode the meaning of words and concepts by semantic relations between them. Hence, the only way to represent a thematic role is a relation between a verb and a noun concept. This relation is semantically characterised by an atomic label. As the wordnet formalism does not provide an apparatus to formally define the semantics of different relation types in general, it does not allow a further formal semantic specification of such a label (i.e. a role type). Instead, the different role labels are characterised by informal definitions and test sentences, entailing the vagueness natural language brings about. Section 2.3.1 will present these definitions.

Jackendoff notes that one important role is not captured by Gruber's work (and earlier work of Jackendoff himself), namely *Patient* (or *Undergoer*), which is intuitively defined as "the affected entity". For example, in

(2.27) Sue hit Fred.

"Fred" is the Patient. There is no role in Gruber's inventory that captures the property of being the affected entity. In particular, "Fred" is not the Theme, because he is not referred to as "thing in motion or being located". Therefore, to properly represent the Patient role, one basic refinement of the role inventory that Jackendoff suggests is its division into two independent tiers, the *thematic tier* and the *action tier*. The thematic tier comprises the roles mentioned so far. The action tier comprises the Patient role and its "counterpart" *Actor*, the "doer of the action". This division is motivated by the fact that it is not unusual for verbal complements to bear two roles at the same time, one from the thematic tier and one from the action tier. The following examples illustrate the two tiers:

(2.28)	Sue	hit	Fred.	
	Theme		Goal	(thematic tier)
	Actor		Patient	(action tier)

(2.29) Pete threw the ball.
Source Theme (thematic tier)
Actor Patient (action tier)

(2.30) Bill entered the room.
Theme Goal (thematic tier)
Actor (action tier)

Analogously to the other roles, we need a representation for Actor and Patient in terms of conceptual structure. Moreover, this representation must be distinct from the representation of the roles of the thematic tier. To achieve this, Jackendoff introduces the conceptual function AFF (affect). This function takes two arguments; the first argument represents the Actor, the second one the Patient. In a conceptual structure, AFF appears in addition to the functions which represent the thematic tier, as in (2.31).[8]

(2.31) Harry gave Sam a book.

$$\begin{bmatrix} \text{CAUSE ([HARRY], [GO}_{\text{Poss}}\text{ ([BOOK], } \begin{bmatrix} \text{FROM ([HARRY])} \\ \text{TO ([SAM])} \end{bmatrix}\text{)]}) \\ \text{AFF}^{+}\text{ ([HARRY], [SAM])} \end{bmatrix}$$

As example (2.30) shows, it can be the case that the Actor or the Patient are not referred to by an utterance. In these cases, the respective argument of AFF is unspecified, such as the Actor argument in (2.32):

(2.32) Sam received a book.

$$\begin{bmatrix} \text{GO}_{\text{Poss}}\text{ ([BOOK], [TO ([SAM])])} \\ \text{AFF}^{+}\text{ (\quad, [SAM])} \end{bmatrix}$$

Animate beings can be volitional or nonvolitional Actors. Some verbs require volitional actors (e.g. "buy" or "look"), some require nonvolitional actors (e.g. "die"), and some are ambiguous in this respect, like "roll":

(2.33) Bill rolled down the hill.

Of course, inanimate things are always nonvolitional:

(2.34) The ball rolled down the hill.

In general, ambiguous cases with animate Actors have a preferred volitional reading.

To distinguish volitional and nonvolitional Actors, Jackendoff elaborates the function AFF by ascribing it the binary feature [±vol] (notated as a subscript of AFF). Thus, the conceptual structure for (2.34) is

[8] In cases like this one, where Harry did something *for* rather than *to* Sam, Jackendoff calls the second argument of AFF *Beneficiary* rather than Patient. Formally, he distinguishes these two roles by an additional feature attached to AFF: AFF^{-} has a Patient, AFF^{+} a Beneficiary as its second argument.

(2.35) $\begin{bmatrix} \text{GO}_{\text{Poss}}\ ([\text{BALL}], [\text{DOWN}\ ([\text{HILL}])]) \\ \text{AFF}_{-\text{vol}}\ ([\text{BALL}]\ ,\quad) \end{bmatrix}$

As the reader may have noted already, the conceptual representation of Agent on the thematic tier (the first argument of CAUSE) and Actor on the action tier (the first argument of AFF) imply a subdivision of the traditional notion of Agent into 'extrinsic instigator' on the one hand and '(volitional or nonvolitional) actor' on the other hand. For example, "Bill" is the extrinsic instigator in (2.36), but not in (2.37):

(2.36) Bill rolled the ball down the hill.

(2.37) Bill rolled down the hill.

Of course, these roles are related; in (2.36), "Bill" acts on the ball, and hence also bears the Actor role. This analysis is an example of how the conceptual representation of roles clarifies their delimitations and their relationships to each other.

The traditional Instrument role is an example for a rather complex structural representation of a thematic role. Jackendoff characterises this role in the following way:

> (1) it plays a role in the means by which the Actor accomplishes the action [...] (2) the Actor acts on the Instrument (3) the Instrument acts on the Patient. (Jackendoff 1990, p. 142)

The formalisation of this analysis is illustrated by the following conceptual structure:

(2.38) Sue hit Fred with the stick.

$$\begin{bmatrix} \text{CAUSE}\ ([\text{SUE}], [\text{INCH}\ ([\text{BE}\ ([\text{STICK}], \text{AT}\ ([\text{FRED}]))])]) \\ \text{AFF}\ ([\text{SUE}], [\text{FRED}]) \\ [\text{BY}\ (\begin{bmatrix} \text{CAUSE}\ ([\text{SUE}], [\text{AFF}\ ([\text{STICK}], [\text{FRED}])]) \\ \text{AFF}\ ([\text{SUE}], [\text{STICK}]) \end{bmatrix})] \end{bmatrix}$$

(INCH is the inchoative function. It refers to an event which terminates in the state which is specified by the argument.) The argument of BY expresses the means by which the action is performed. STICK is at the same time Actor (with FRED as Patient) and Patient (with SUE as Actor).

While a decompositional analysis of roles and their interdependencies as illustrated above is not feasible in the wordnet framework, it would in principle be possible to model the distinction between action tier and thematic tier. To this end, one would have to define the role relation types in a way that reflects this distinction, by employing labels like ACTION_ACTOR and ACTION_PATIENT on the one hand and THEMATIC_AGENT, THEMATIC_SOURCE, THEMATIC_GOAL etc. on the other hand. Whether it would make sense to model this distinction without the representational apparatus of conceptual semantics employed by Jackendoff is a different question, which pertains to the general issue of determining the inventory of thematic roles and their definitions. This issue goes beyond the scope of this thesis.

Jackendoff extensively addresses the problem of linking the syntactic complements of a verb to its semantic roles. Since the semantic roles of a verb can generally be realised by a number of different syntactic patterns (cf. examples (2.2)–(2.4) on page 19), linking is a nontrivial task. One way to specify such a mapping is to stipulate it for each verb in its lexical entry by co-indexing the syntactic constituents which the verb subcategorises for and the corresponding constituents of the verb's conceptual structure. The lexical entry in (2.25) on page 24 is an example for this approach, which, as we have seen in section 1.1.2 (examples (1.10) and (1.11)) is also pursued in NLP systems. However, this method provides an idiosyncratic mapping for each verb, and, even worse, for each meaning of a verb. Thus, any generalisation is missed. Many researchers have noted that the mapping between complements and roles follows certain general principles. (For example, if an Actor is mentioned, it is realised as subject.) Consequently, much effort has been spent to capture and integrate these generalisations in a *linking theory* (cf. (Jackendoff 1990, p. 46) for a short survey of relevant literature).

Jackendoff also presents a "preliminary exposition" (Jackendoff 1990, p. 282) of such a theory within conceptual semantics. Among the strategies proposed in the literature, he basically adopts the approach of *hierarchical argument linking*. This approach assumes an ordered list of thematic roles (*thematic hierarchy*) as well as an ordered list of syntactic relations (*syntactic hierarchy*). Based on these lists, the mapping is done in the following way:

> Following the thematic hierarchy, order the θ-roles in the lexical conceptual structure (LCS) of a verb V from first to nth. To derive the syntactic argument structure of V, map this ordering of θ-roles into the first to nth roles in the syntactic hierarchy. (Jackendoff 1990, p. 246).[9]

Jackendoff states the following thematic hierarchy

(2.39) *Thematic hierarchy*

a. Actor
b. Patient / Beneficiary
c. Theme
d. Location, Source, Goal

and the following syntactic hierarchy

(2.40) *Syntactic hierarchy*

a. [$_S$ NP ...] (Subject)
b. [$_{VP}$ V NP ...] (1st Object)
c. [$_{VP}$ V ... NP ...] (2nd Object)

[9] As noted above, it is common that one conceptual entity occurs in multiple roles (e.g., the Actor of "throw" is also the Source). Such an entity is referred to in multiple substructures of the verb's LCS; formally expressed, these substructures are *bound* (e.g., in the LCS of "throw", the first argument of AFF and the argument of FROM are bound, i.e. they refer to the same entity). In these cases, the *dominant* role, i.e. the role which is highest in the thematic hierarchy, is crucial for the linking process.

A simple example: In

(2.41) The door opened.

we only have a Theme on the one hand and a subject on the other hand, so that mapping is trivial. However, in

(2.42) Bill opened the door.

we have the roles Actor and Theme and the syntactic relations subject and object. In this case, it follows from the linking rule and the hierarchies stated above that the Actor is linked to (e.g. realised as) the subject and the Theme to the object.

Assuming general principles for linking, one can abandon co-indexation from the lexical entries. However, it is still necessary to mark *which* subconcepts of a verb's LCS (lexical conceptual structure) are to be linked to its syntactic complements. In contrast to indexes, these markers do not have to specify *to which* complement the respective substructures have to be mapped, since this is determined by the linking principles. Thus, it is sufficient to introduce a single "marker of argumenthood" (A) as a substitute for the indexes in the LCS. In this way, the LCS in the entry for "run" ((2.25) on page 24) has to look like this:

(2.43) [Event GO ([Thing]A, [Path]A)]

The necessity for A-marking can be demonstrated by comparing (2.43) with the more complex LCS for the verb "enter":

(2.44) [Event GO ([Thing]A, [Path TO ([Place IN([Thing]A)])])]

The meaning of "enter" incorporates the TO and the IN function. These functions do not have to be expressed by the preposition "into" (compare the lexical entry (2.26) on page 24 where TO and IN are specified). Thus, the complement of "enter" that represents the Goal does not express a Path, but a Thing which constitutes the Place at the end of the Path to the Goal, e.g. "the room". Thus, it is possible to say

(2.45) Bill entered the room.

while

(2.46) *Bill ran the room.

is odd, since the LCS of "run" requires a complement which expresses a Path, not a Thing.

Note that the syntactic hierarchy (2.40) only contains NP complements. Other constituent types (PPs, APs, and Ss) are not captured by the linking rule mentioned above. Jackendoff states that such complements are "freely" linked to A-marked conceptual constituents of the verb's LCS. The only condition that must be met is that the conceptual structure corresponding to the complement has to be consistent with the A-marked conceptual constituent to which that complement is linked. Thus, it is not allowed to link an *into*-PP to a FROM constituent, since the conceptual structure for an *into*-PP has TO as its outermost function.

In other words, the linking mechanism sketched above only accounts for NP complements. The rationale behind this is that "NP arguments, unlike others, give no intrinsic hints as to their roles" (Jackendoff 1990, p. 267) and thus require stricter principles for linking, whereas other complements more likely provide such "intrinsic hints" (e.g., the TO function in the LCS of "into" corresponds to the Goal). Furthermore, it is not necessary to account for the relative positions of non-NP complements in a sentence by linking hierarchies, because these positions are entirely constrained by syntactic regularities (e.g., complements appear in the order NP, AP, PP, S; if several PPs occur, then their order is free).

As noted several times, developing a linking strategy is one of the central issues of this thesis. The strategy for linking I will present in chapter 6 fundamentally differs from the approach of hierarchical argument linking adopted by Jackendoff. This approach requires that both the syntactic and the semantic arguments of a verb are known. The linking principles stated above establish an appropriate mapping between them. In contrast, the data I use only contain information about the syntactic arguments of a verb; information about their semantic arguments is not given. For this reason, my linking strategy has to apply other means (semantic clustering techniques combined with general heuristics) to identify semantic roles.

2.1.3 Dowty's Proto-Roles

To motivate his analysis of thematic roles, Dowty (cf. (Dowty 1991, p. 547–552)) summarises the linguistic research in this field and states that "there is in fact a notable absence of consensus about what thematic roles are" (Dowty 1991, p. 547). We have already noted this problem, which is the actual reason for the difficulty to determine a set of thematic roles, in particular the granularity of this set and the exact boundaries between its members. One way to circumvent these problems, which was argued for in the literature (cf. (Dowty 1991, p. 547–55)), is to assume *individual thematic roles* for each verb, rather than verb-independent thematic role types. For example, instead of an Agent role, there would be a "hitter role" associated with the verb "hit", a "killer role" with "kill", a "builder role" with "build", etc. Dowty rejects this idea, because it misses syntactic generalisations related to thematic roles (e.g. a general linking theory).

Dowty argues that the investigation of thematic roles is less complex if one, instead of trying to cope with all issues connected with thematic roles at the same time, concentrates on a single linguistic problem for which thematic roles are significant. In a second step, one can explore whether the results for one domain are applicable to other domains. Following this strategy, (Dowty 1991) focuses on the problem of *argument selection*. (At the end of the article, he shortly outlines how his results could be transferred to the domains of language acquisition and the unaccusative-unergative distinction.) Argument selection deals with the principles that determine which semantic arguments of a verb are expressed by which grammatical relation (subject, object). Fillmore's subject selection principle (Fillmore 1968) and his saliency hierarchy (Fillmore 1977) as well as Jackendoff's hierarchies for

argument linking fall within this domain.

Dowty argues for treating thematic roles not as discrete categories, but as prototypical concepts, so that "degreed membership" of arguments in different role types is possible. For his theory of argument selection, he defines two *proto-roles*: Proto-Agent (P-Agent) and Proto-Patient (P-Patient) (cf. (Dowty 1991, p. 572)). Each of these proto-roles are characterised by a set of entailments (properties):

(2.47) Proto-Agent

a. volitional involvement in the event or state
b. sentience (and/or perception)
c. causing an event or change of state in another participant
d. movement (relative to the position of another participant)
(e. exists independently of the event named by the verb)

(2.48) Proto-Patient

a. undergoes change of state
b. incremental theme
c. causally effected by another participant
d. stationary relative to movement of another participant
(e. does not exist independently of the event, or not at all)

Usually, for a verb's argument, several of these entailments co-occur. However, they are semantically independent. To illustrate the individual entailments and their independence, Dowty provides the examples in (2.49) and (2.50). (2.49) lists cases where just one of the entailments holds for the P-Agent (at the subject position):

(2.49)

a. *Volition:* John is being polite to Bill / is ignoring Mary.
b. *Sentience/Perception:* John knows / believes / is disappointed at the statement. John sees / fears Mary.
c. *Causation:* His loneliness causes his unhappiness. Teenage unemployment causes delinquency.
d. *Movement:* The rolling tumbleweed passed the rock. The bullet overtook the arrow. Water filled the boat. He accidently fell.
e. *Independent existence:* John needs a car.

(2.50) lists analogous cases for the P-Patient (at the object position):

(2.50)

a. *Change of state:* John made a mistake *(Coming into being, therefore also e. below)*. John moved the rock *(indefinite change of position)*. John erased the error *(ceasing to exist)*.

b. *Incremental Theme:* John crossed the driveway / filled the glass with water *(also d.)*.

c. *Causally affected:* Smoking causes cancer.

d. *Stationary relative to another participant:* The bullet entered the target / overtook the arrow.

e. *Existence not independent of event:* John built a house / erased the error *(Coming into and out of existence; not independent of a.)*. The situation constitutes a major dilemma for us. John needs a car / seeks a unicorn / lacks enough money to buy it *(de dicto objects: no existence)*.

The notion *Incremental theme* needs explanation. It refers to the Theme of a telic predicate and is motivated "by the principle that *the meaning of a telic predicate is a homomorphism from its (structured) Theme argument denotations into a (structured) domain of events*" (Dowty 1991, p. 567). This homomorphism maps the part-of relation in the domain of the Theme into the part-of relation in the domain of telic events. Consider, for example, the event denoted by "John filled the glass with water". The amount of water which actually is in the glass corresponds to the 'aspect' of the filling event: If the glass is empty, the event has not yet begun; if it is partially filled, then the event is partially done; if it is full, then the event is completed.

With the help of the P-Agent and the P-Patient roles, Dowty states the *argument selection principle*:

> In predicates with grammatical subject and object, the argument for which the predicate entails the greatest number of Proto-Agent properties will be lexicalised as the subject of the predicate; the argument having the greatest number of Proto-Patient entailments will be lexicalised as direct object. (Dowty 1991, p. 576)

According to this principle, there are more or less prototypical cases for argument selection. For predicates like "build", "murder", "eat", or "wash", the subjects have several P-Agent entailments (volition, sentience, causation, movement) and no P-Patient entailments, and the objects have no properties of P-Agent, but several properties of P-Patient (change, causally affected, and most of them incremental theme, stationary, dependent existence). On the other end of the scale, there are "lexical doublets" like "buy" and "sell". In the buy/sell event, exactly the same entailments hold for the buyer and the seller. The argument selection principle cannot decide which one of them should be realised as subject and which one as object. Thus, it is consistent with that principle that both possibilities are lexicalised (by the verbs "buy" and "sell", respectively).

A more complex case is the class of *spray/load* verbs, illustrated in (2.20) and (2.21) on page 22. Dowty explains this alternation in terms of P-Patient entailments. In

(2.51) Mary loaded the hay onto the truck.

both the hay and the truck undergo a change of state. Additionally, the hay is Incremental Theme, since the sentence implies that there was a certain amount of hay that is completely loaded onto the

truck when the event is completed. However, it is not implied that the truck is completely loaded when the event is completed, so the truck is not an Incremental Theme. Thus, "the hay" has more P-Patient entailments than "the truck", and is placed at the object position. The reverse is true for

(2.52) Mary loaded the truck with (the) hay.

Here, the truck is Incremental Theme, since the sentence implies that it is completely loaded when the loading event has finished. The hay is no Incremental Theme, since it is not necessarily implied that the total amount of hay is loaded. Thus, "the truck" is realised as object here.[10]

Dowty notes that the traditional role types can be reformulated in terms of proto-role entailments. Moreover, different conceptions of particular role types can be captured by different combinations of entailments:

> *Agent* is volition + causation + sentience + movement, or in some usages just volition + causation, or just volition [...] or, according to the ordinary language sense of 'agent', causation alone. *Experiencer* is sentience without volition or causation. *Instrument* is causation + movement without volition or sentience. *Theme* (excepting Gruber's stative theme) is most typically change + Incremental Theme + dependent-existence + causally-affected, but causally-affected is sometimes absent (*Patient* can be distinguished from broader Theme by this entailment [...]" (Dowty 1991, p. 577).

Thus, the system of proto-role entailments explicates similarities and differences between roles and is more flexible with respect to the granularity of the role inventory, since it allows broader or finer differentiation by providing different dimensions of distinction.

Moreover, characterising role types in terms of entailments together with the argument selection principle yields rules for linking which are similar to those sketched in the two previous sections. For example, an Agent generally appears as the subject of a sentence, since it has the most P-Agent entailments. In cases where no Agent is present, Instruments or Experiencers are realised as subject rather than Patients, because the former roles have more P-Agent and less P-Patient entailments than the latter. Furthermore, Patients outrank Sources and Goals for objecthood, since Patients have several P-Patient properties, whereas Sources and Goals only have one, namely stationary. Dowty summarises rules of this kind by a role hierarchy which is similar to the hierarchy in (2.39)

(2.53) $$\text{Agent} > \left\{ \begin{array}{l} \text{Instrument} \\ \text{Experiencer} \end{array} \right\} > \text{Patient} > \left\{ \begin{array}{l} \text{Source} \\ \text{Goal} \end{array} \right\}$$

and the following additional rankings (Arg refers to an argument for which no proto-role entailment holds):

(2.54)
$$\begin{array}{lcl} \text{causing event} & > & \text{caused event} \\ \text{moving argument} & > & \text{Source, Goal, Arg} \\ \text{Experiencer} & > & \text{Arg} \end{array}$$

[10] Actually, the argument is much more subtle, e.g. concerning the determiner of "hay". See (Dowty 1991, p. 587–592) for details.

Since information about entailments of proto-roles as specified in (2.47) and (2.48) are not provided by the data I employ as input for my learning algorithm, Dowty's approach in a strict sense is not applicable for the task of this thesis. However, for my linking strategy, I employ a semantic clustering technique which can be viewed as a probabilistic implementation of Dowty's theory (cf. section 6.2, in particular section 6.2.2). In this view, a verb–noun cluster derived by this technique is analogous to a proto-role entailment. (Of course, which kind of entailment a cluster represents is not determined by the clustering method.) The thematic role underlying a syntactic argument of a verb is determined by its pattern of occurrence in the different clusters.[11]

2.2 Selectional Restrictions

2.2.1 Traditional Treatment of Selectional Restrictions

The notion of selectional restrictions was originally introduced within the transformational grammar paradigm in the sixties of the last century. This section summarises the main considerations and ideas concerning selectional restrictions that emerged in that period.

Chomsky

A particularly controversial issue in the discussion of selectional restrictions was whether they are a syntactic or a semantic phenomenon. Chomsky (cf. (Chomsky 1965)) models selectional restrictions as part of the syntactic component. In his theory, nouns are characterised by features like [+Animate], [+Human], or [−Abstract]. The encoding of selectional constraints refers to such features. Chomsky claims that they are syntactic features, because they play a role in purely syntactic rules. For example, the sentence

(2.55) *The book who you read was a best seller.

violates rules which are syntactic in nature and refer to the feature [Human].

A lexicon entry for a word contains, apart from its phonological form, a so-called *complex symbol*, i.e. a bundle of syntactic features, as illustrated by the following entries for "sincerity" and "boy", respectively:

(2.56) (*sincerity*, [+N, +Det—, −Count, +Abstract,...])

(2.57) (*boy*, [+N, +Det—, +Count, +Animate, +Human,...])

As mentioned, features like [−Count], [+Abstract], [+Animate], and [+Human] above are crucial for characterising the selectional restrictions that a verb imposes on its arguments. For example, "frighten" selects an abstract subject and an animate object. The lexical entry for this verb is

[11] The clustering technique yields soft clusters, i.e. an argument is member of different clusters to different degrees.

(2.58) (*frighten*, [+V, +——NP, +[+Abstract] Aux——Det [+Animate], +Object-deletion,...])

In this entry, selectional restrictions are expressed by the feature

(2.59) [+[+Abstract] Aux——Det [+Animate]]

Technically, this feature encodes conditions for the deep structure context into which the verb "frighten" may be inserted. It states that if "frighten" occurs in a deep structure, then it must be immediately preceded by a category with the feature [+Abstract] and an auxiliary, and it must be immediately followed by a determiner and a category specified as [+Animate]. The grammar rules in Chomsky's base component (cf. (Chomsky 1965, p. 106–107)) provide the features [±Animate] and [±Abstract] for nouns. Furthermore, they introduce an obligatory auxiliary category (Aux) preceding the VP. The NP which precedes Aux is the subject, the NP which follows the main verb is the object of the sentence. Thus, the restriction [+Abstract] refers to the head of the subject NP, while the restriction [+Animate] refers to the head of the object NP.

Analogously, selectional restrictions which an adjective imposes on the noun it modifies are encoded in the lexicon entry of the adjective. For example, the entry for "green" contains a feature which specifies that the modified noun is [−Abstract].

Chomsky considers the possibility to encode selectional restrictions of a verb separately for each of its arguments instead of stating constraints on the complete set of arguments. With this approach, the single specification (2.59) in the lexicon entry for "frighten" would be replaced by the two features

(2.60) [+[+Abstract] Aux——, +——Det [+Animate]]

i.e., the selectional restrictions for the subject and the object would be encoded independently from each other. However, there are cases in which the selectional restrictions for the different arguments of a verb are not independent. As an example, Chomsky mentions the following contexts:

(2.61)

a. He —— the platoon
b. His decision to resign his commission —— the platoon
c. His decision to resign his commission —— our respect

The verb "command" can be inserted in a. and c. but not in b. "command" can have an animate or an abstract subject, as well as an animate or an abstract object. If we model this information as in (2.60), i.e. specify the restrictions

(2.62) [+[+Animate] Aux——, +[+Abstract] Aux——]

for the subject and

(2.63) [+——Det [+Animate], +——Det [+Abstract]]

for the object, then we do not express the restriction that either both arguments have to be animate, or both have to be abstract. To account for this constraint, we have to encode restrictions for the subject and the object together:

(2.64) [+[+Animate] Aux——Det [+Animate], +[+Abstract] Aux——Det [+Abstract]]

This way of encoding captures dependencies between selectional restrictions of different arguments.

As noted already in section 2.1.1, the approach proposed in this thesis and comparable approaches for acquiring selectional preferences (cf. chapter 3) do not model such dependencies between different arguments of a verb. The reasons for this decision are explained in section 3.2.

Generally, the violation of selectional restrictions results in an unnatural, anomalous sentence like

(2.65) *John frightened sincerity.

or Chomsky's famous example

(2.66) *Colorless green ideas sleep furiously.

However, in certain environments (negation and/or embedded constructions constituting a certain "meta level"), such a violation does not yield unnaturalness:

(2.67) It is nonsense to speak of frightening sincerity.

(2.68) One can(not) frighten sincerity.

However, this phenomenon is not limited to selectional restrictions. The sentence

(2.69) One can(not) elapse a book.

is not deviant either, although it violates rules of strict subcategorisation, since "elapse" does not subcategorise for an object.

For current approaches which acquire selectional restrictions from corpus data, such examples would introduce undesired noise. As the kind of "meta context" they are embedded in cannot easily be recognised automatically, these examples cannot be eliminated from the data employed by such approaches. Thus, the effectiveness of these methods rests on the reasonable assumption that such examples occur only infrequently in "real" text.

Although Chomsky models selectional restrictions as part of the syntactic component, he does not reject the alternative possibility to treat them as part of the semantic component. He states that the final decision on this issue requires deeper knowledge about syntax and semantics in general (cf. (Chomsky 1965, p. 159–160).

Katz/Fodor

In contrast to (Chomsky 1965), (Katz & Fodor 1964) treat selectional restrictions as a semantic phenomenon. In their framework, selectional restrictions play a crucial role for resolving ambiguities during the semantic interpretation of syntactic structures. Recall that the suitability of selectional constraints for ambiguity resolution has been illustrated in section 1.1.1.

The different senses of a word are encoded in its lexicon entry by semantic features. For example, (Katz & Fodor 1964) mention the following senses of the word "bachelor":

(2.70)

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

The features in parentheses are called *semantic markers*, whereas the expressions in square brackets are called *distinguishers*. Semantic markers are "the systematic features of the semantic structure of the language" (Katz & Fodor 1964, p. 500). These are the features which are necessary to systematically delimit the different word senses (while the distinguishers provide idiosyncratic aspects of each sense). Thus, semantic markers contain the information which is needed for semantic disambiguation. They are employed to encode selectional restrictions.

As we saw in the previous subsection, one can argue that features like (Male), (Female), (Human), (Abstract) etc. are syntactic features, because they are used to model purely syntactic phenomena. However, (Katz & Fodor 1964) claim that in cases where a feature is apparently involved in both syntactic and semantic regularities, then there are in fact two features with the same name (or similar names) but different domains. To support this view, Katz and Fodor mention several cases in which a syntactic feature does not completely coincide with the corresponding semantic feature. For instance, "baby" has the semantic marker (Human) which implies either the marker (Male) or (Female). However, "baby" is neither syntactically classified as masculine nor as feminine, but as neuter:

(2.71) The baby lost its rattle.

Selectional constraints which a verb imposes on its arguments are encoded in its lexicon entry by Boolean combinations of semantic markers. E.g., (one sense of) the verb "hit" has the following selectional restrictions:

(2.72) $<$SUBJECT: (Human) $\vee$ (Higher Animal),
OBJECT: (Physical Object),
INSTRUMENTAL: (Physical Object)$>$

Likewise, the lexicon entry of an adjective includes (a Boolean expression of) semantic markers encoding the restrictions on the noun it modifies. Thus, "colorful" (for its literal meaning) requires

that the modified noun is marked as (Physical Object).[12]

In principle, the application of selectional restrictions for disambiguation in the account of Katz/Fodor works analogously as in NLP systems discussed in section 1.1.2. The selectional restrictions encoded in the lexicon are checked during the semantic interpretation of a sentence. This interpretation is carried out by compositional *projection rules*, which act on syntactic structures (generated by a transformational grammar). The projection rules determine how the semantics of a constituent is composed (*amalgamated*, as Katz and Fodor call it) from the semantics of its sub-constituents. Selectional restrictions constrain the amalgamation of a predicate and its arguments: The semantic representations of a predicate constituent and an argument constituent are combined only if the argument representation meets the selectional conditions specified in the predicate representation.

For illustration, Katz and Fodor discuss the sentence

(2.73) The man hits the colorful ball.

In the first step, the semantics of "colorful ball" is amalgamated from the semantics of "colorful" and "ball". The abovementioned selectional restrictions of "colorful" require that "ball" is marked as (Physical Object). This is the case for the 'round object' sense of "ball" (as opposed to the 'dancing event' sense, which is marked as (Social Activity)), so the semantic representation of this sense is amalgamated with the semantic representation of "colorful". In a later step, the semantics of "the colorful ball" is amalgamated with the semantics of "hit". The projection rule which is applied to the corresponding syntactic sub-structure recognises "the colorful ball" as the object of "hit" and checks whether the verb's restrictions for its object—(Physical Object)—are met. Again, this is the case so that the two constituents can be semantically combined.

The treatment of the ambiguity of "ball" just sketched shows how selectional restrictions are employed for disambiguation. In general, a constituent that comprises ambiguous words can have several semantic representations, namely all those combinations of the semantic representations of its sub-constituents which are not inhibited by selectional restrictions. For example, apart from its literal sense, "colorful" has the metaphorical reading 'of distinctive vividness'. Therefore, the NP "the colorful ball" can have the meanings (among others) 'round object exhibiting a variety of bright colors' or 'dancing event of distinctive vividness'. However, the meaning 'round object of distinctive vividness' is excluded by the selectional restrictions of the 'vividness' sense of "colorful", which is not applicable to physical objects. The combination with "hit" further disambiguates "the colorful ball": The selectional restrictions of "hit" require that its object is marked as (Physical Object). Thus, senses involving 'dancing event' are rejected.

In the same way, selectional restrictions can resolve syntactic ambiguities. If the grammar generates several syntactic structures for a sentence, then semantic interpretation is applied to each of these structures, and it may be that some of these syntactic structures do not receive a semantic representation because selectional restrictions do not permit this. In this way, structural ambiguities are resolved. If there are still multiple structures which are assigned a semantic interpretation, then the sentence is syntactically ambiguous. If, however, none of the syntactic structures receives a semantic representation, then the sentence is semantically anomalous.

[12] In addition to selectional restrictions, lexical verb or adjective entries, like nominal entries, contain markers and distinguishers which characterise their meaning. E.g., "hit" is marked as (Action), "colorful" as (Color).

McCawley

Unlike (Chomsky 1965) and (Katz & Fodor 1964), McCawley (cf. (McCawley 1968, p. 125–136)) does not propose a formal theory of selectional restrictions. However, he makes some important statements about them. He provides convincing evidence for regarding selectional restrictions as a semantic rather than a syntactic phenomenon. He claims that selectional restrictions refer to semantic, not to syntactic properties. For example, there is no English verb which requires that its subject is syntactically classified as feminine, i.e. pronominalised to "she" (which is the case for women, ships, and countries). McCawley also discusses some apparent counter-examples. The verb "name" seems to restrict its second object to proper nouns:

(2.74) They named their son John.

(2.75) *They named their son that boy.

However, the sentence

(2.76) They named their son something outlandish.

shows that this is not the case. Likewise, "count" seems to require a plural object:

(2.77) I counted the boys.

(2.78) *I counted the boy.

But there are possible singular objects for this verb:

(2.79) I counted the crowd.

Thus, the restriction here is not that the object has to be *syntactically* classified as plural, but that it denotes a set of entities (as opposed to a singular entity). This is a *semantic* property.

These examples suggest that selectional restrictions only refer to semantic information. McCawley argues that if selectional restrictions are a semantic phenomenon, then it is appropriate to capture them by the semantic component of a linguistic theory, as in (Katz & Fodor 1964), rather than by the syntactic component, as in (Chomsky 1965). This view has been generally accepted as well in computational linguistics so that, as illustrated in chapter 1, selectional constraints are encoded as lexical-semantic information in NLP systems and resources.

Now the question arises whether there is *a limited set* of semantic features which is sufficient to encode selectional restrictions appropriately, or whether *any* semantic information may play be relevant for them. McCawley argues for the latter:

> "... on any page of a large dictionary one finds words with incredibly specific selectional restrictions, involving an apparently unlimited range of semantic properties. For example, the verb *diagonalise* requires as its object a noun phrase denoting a matrix (in the mathematical sense), the adjective *benign* in the sense 'noncancerous' requires a subject denoting a tumour, and the verb *devein* as used in cookery requires an object denoting a shrimp or prawn." (McCawley 1968, p. 134)

This finding suggests the use of large-scale semantic hierarchies such as WordNet rather than small inventories of semantic primitives (as in the NLP systems described in section 1.1.2) to encode selectional restrictions. Although no lexical-semantic resource will be "complete" in the sense that it captures any imaginable semantic concept, a high coverage of concepts is desirable to be able to capture very specific selectional restrictions as appropriately as possible. Furthermore, McCawley's examples show that the task of finding the appropriate generalisation level for representing the selectional constraints of a particular predicate, a core issue of this thesis, is all but trivial: Some predicates (such as "count") impose rather general restrictions on their arguments, others (such as "devein") rather specific constraints.

Apart from the treatment of selectional restrictions as a semantic phenomenon, there is an additional reason for favouring the approach of Katz and Fodor over that of Chomsky. While Chomsky regards selectional restrictions as relating two lexical items (e.g. a verb and the head noun of its subject), Katz and Fodor model them as relating a lexical item and a constituent (e.g. a verb and its subject NP). As McCawley points out, only the latter approach accounts for the fact that selectional restrictions can be violated not only by the head of a complement, but also by a modifier of that head. The sentence

(2.80) *My buxom neighbor is the father of two.

violates selectional restrictions in the same way as the sentence

(2.81) *My sister is the father of two.

This violation is caused by the modifier in "my buxom neighbor", since

(2.82) My neighbour is the father of two.

does not exhibit a violation.

Despite these facts, I will adopt the decision which generally has been made for current approaches of learning selectional preferences from corpora. These approaches only take the head noun of a complement into account. There are practical reasons for this simplification; cf. section 3.2 for further explanations.

McCawley criticises one important aspect of Katz' and Fodor's approach, namely that selectional restrictions are the only means for resolving semantic ambiguities. He rather claims that several factors contribute to disambiguation, world knowledge in particular. In the sentence

(2.83) My aunt is a bachelor.

Katz' and Fodor's theory would disambiguate "bachelor" as 'who has the first or lowest academic degree'. The sense 'who has never married' would be rejected, because this sense is marked as (Male). However, in many contexts, e.g. when talking about a spinster, the latter meaning is much more obvious than the former.

Moreover, McCawley criticises that Katz and Fodor do not assign semantic representations to constituents that violate selectional restrictions. If semantically anomalous constituents were meaningless, then the sentences

(2.84) He says that he smells itchy.

(2.85) He says that he poured his mother into an inkwell.

(2.86) He says that his toenail sings five-part madrigals.

would have to be regarded as synonymous. McCawley concludes

> "...that the violation of selectional restrictions is only one of many grounds on which one could reject a reading as not being what the speaker intended and that it moreover does not hold any privileged position among the various criteria for deciding what someone meant." (McCawley 1968, p. 130)

To a certain extent, these considerations can be accounted for by modelling selectional restrictions as preferences rather than hard constraints. In section 2.3.2, I will argue for this option.

2.2.2 Selectional Restrictions and Thematic Roles

In this section, I turn to the question of how thematic roles and selectional restrictions are related.

As we saw in section 2.2.1, selectional restrictions of verbs are traditionally treated as constraints on grammatical relations, i.e. subjects, objects etc. However, we also saw that selectional restrictions are semantic, not syntactic constraints. As the examples (2.2)–(2.4) on page 19 illustrate, the semantic roles of a verb can be realised by different patterns of syntactic complements. Accordingly, a certain grammatical relation can express different semantic roles of a given verb. Therefore, it seems more appropriate to model selectional restrictions as referring not to surface syntactic relations, but to the underlying semantic roles. Recall that this has been done in NLP systems such as Mikrokosmos (cf. the lexical entries (1.10) and (1.11) in section 1.1.2).

It is important to note that this view is no contradiction to the approaches described in section 2.2.1. In transformational grammar, selectional restrictions refer to grammatical relations in the deep structure. Different syntactic realisations of verbal arguments are generated by transformations (e.g. the passive transformation) at a later stage of sentence generation. Thus, in the deep structure, a particular grammatical relation corresponds to a particular semantic argument for a given verb.[13] Recall that

[13]The theory of (Katz & Fodor 1964) is slightly more sophisticated in this point. In principle, "projection rules can take account of information about the transformational history of a sentence" (Katz & Fodor 1964, p. 505), i.e. are not necessarily acting on deep structures. However, in cases where a sentence has undergone optional transformations which account for syntactic alternations, the projection rules are applied to the corresponding "kernel sentence" which was generated without these optional transformations.

in Fillmore's case grammar (according to (Fillmore 1968)), the deep cases are major constituents of the deep structure (cf. example (2.11) on page 20). These constituents represent the arguments of the verb, i.e. are just those constituents to which selectional restrictions are applied.

The literature about thematic roles sketched in section 2.1 provides further support for the view that selectional restrictions refer to semantic roles. As noted in section 2.1.1, (Fillmore 1968) points out that different deep cases have different semantic restrictions: Agent and Dative are (typically) animate, Instrument is inanimate, Locative denotes some location. These constraints are of the same sort as selectional restrictions.

(Jackendoff 1990, p. 50–55) provides a principled account of selectional restrictions within conceptual semantics. Selectional constraints of a verb are modelled within its lexical conceptual structure (LCS), as elaborations of the sub-structures which represent the verb's arguments. In other words, selectional restrictions are connected with thematic roles. For example, "drink" selects liquids for its Theme. The LCS of this verb (specified in its lexical entry) approximately looks like this:

(2.87) $[_{\text{Event}}$ CAUSE $([_{\text{Thing}}\ ]^{\alpha}{}_{\text{A}}, [_{\text{Event}}$ GO $([_{\text{Thing}}$ LIQUID$]_{\text{A}},$
$[_{\text{Path}}$ TO $([_{\text{Place}}$ IN $([$MOUTH OF $([\alpha])])])])])]$[14]

Here, the Theme (the first argument of GO) is specified by the concept LIQUID. In principle, any conceptual (sub-)structure that is suitable in the respective environment can be used to characterise selectional restrictions. (This corresponds to McCawley's observation that any semantic information can be relevant for selection.)

Jackendoff emphasises that selectional restrictions are not necessarily implied by the action denoted by the verb. For example, the action 'pour down one's throat' (i.e. the action denoted by "drink") is not restricted to liquid Themes, since one can pour down powder as well. However, "drink" provides the additional information that what you pour down is liquid. Likewise, "pay" refers to the action "give in exchange for something", but additionally provides the information that what you give is money, which cannot be concluded from the exchange action itself.

The formulation "provides the information" instead of "imposes the restriction" in the previous paragraph indicates that modelling selectional restrictions by conceptual structures yields an understanding of this term that differs significantly from the traditional view: Jackendoff regards selectional restrictions not as constraints that a verb imposes on its argument, but as information that a verb supplies about its argument, as an integral part of the verb's meaning. Evidence for this view is provided by cases where the argument is pronominalised or omitted completely:

(2.88) Harry drank it.

(2.89) Harry drank (again).

(2.90) Bill paid Harry a lot.

(2.91) Bill paid.

[14]The notation involving α means that the argument of MOUTH OF is bound to (i.e. identical with) the first argument of CAUSE.

In these cases, the object itself is underspecified. But we still know that Harry drank some liquid and Bill paid some amount of money. This information is supplied by the verb. To obtain the complete information which is given for an argument, the information expressed by the verb for that argument and the information provided by the corresponding complement is merged. (Jackendoff calls this operation *argument fusion*.)[15] For example, in sentence (2.90), the Theme is "a lot of money". The information "a lot" is given by the object, the information "money" by the verb. In cases of omitted arguments, the verb is the only supplier of information about the argument.

If the verb and a complement supply conflicting information about an argument, then there is a violation of selectional restrictions. In

(2.92) Harry drank powder.

the conceptual structure of "powder", which contains the feature SOLID, is fused with the selectional restriction LIQUID, resulting in a conflict. In contrast,

(2.93) Harry drank sincerity.

exhibits a violation already at the level of major category: "drink" selects a Theme of the category Thing, whereas "sincerity" has the major category Property.

Incorporated arguments are modelled in the same way as selectional restrictions. For instance, the verb "butter" as in

(2.94) Harry buttered the bread.

has the following LCS:

(2.95) [Event CAUSE ([Thing]A, [Event GO ([Thing BUTTER], [Path TO ([Place ON ([Thing]A)])])])]

Here, the incorporated Theme is further specified by BUTTER. This concept encodes the information which the verb provides for its Theme. The only difference between an argument including a selectional restriction and an incorporated argument is that the latter is not A-marked, so that it will not be linked to—and hence not expressed by—a syntactic complement of the verb. Thus, the representation of selectional restrictions does not differ from the representation of other parts of a verb's meaning. Therefore, Jackendoff concludes that "the notion of selectional restrictions can be dropped altogether from linguistic theory except as a convenient name for the effects of Argument Fusion" (Jackendoff 1990, p. 55).

Anyway, this conclusion does not prevent Jackendoff from using this "convenient name" as a matter of course, which indicates that the notion of selectional restrictions nonetheless is of significance. In fact, selectional restrictions play an important role in Jackendoff's linking theory. Recall that in this theory non-NP complements are linked "freely"—without employing specific linking principles—to

[15] In NLP, this merging of information would be implemented as unification.

A-marked constituents of the verb's LCS. Linking implies argument fusion: the conceptual structure of an argument and the A-marked constituent which is linked to that argument are fused. This fusion is possible only if both are semantically compatible, i.e. if the argument's conceptual structure is consistent with the conceptual elaboration of the A-marked constituent. This elaboration is just the information which constitutes the verb's selectional restrictions for that argument. This ensures that e.g. a *to*-PP complement of "give" is linked to the GOAL, because the GOAL argument is specified as $[\text{TO}([\ ])]_{\text{A}}$ in the LCS of "give".

Jackendoff regards selectional restrictions as important only for linking non-NP complements, since NP complements are captured by the linking hierarchies. However, it seems to me that there are cases of NP linking in which selectional restrictions are necessary as well. In the example Jackendoff discusses to demonstrate NP linking ("Bill opened the door." vs. "The door opened."), the two variants differ in the number of arguments (two arguments for the causative, one for the inchoative sense of "open"). Correspondingly, Jackendoff provides two different LCSs for this verb, one including Agent and Theme, and one including only Theme. Here, the correct assignment of the two LCSs to the corresponding syntactic patterns follows from the linking mechanism: Since the mapping between syntactic complements and A-marked constituents has to be bijective, the numbers of arguments have to match. However, consider cases like

(2.96) Mother is cooking.

(2.97) The potatoes are cooking.

(cf. (Fillmore 1968)). Here, the two variants have the same number of arguments. Moreover, they share the same syntactic pattern. It is ambiguous whether the subject is Actor or Patient. In Jackendoff's framework, this ambiguity would be captured by two different LCSs (or a disjunctive LCS) where *either* the Actor *or* the Patient is A-marked.[16] Since (2.96) and (2.97) are syntactically identical, only semantic information, i.e. selectional restrictions, can help to decide which LCS is appropriate for which alternative. The restriction that "cook" requires an animate Actor rejects the possibility that the potatoes are the Actors in (2.97); they must be Patient. (2.96) is really ambiguous, although the reading where "mother" is the Actor is highly preferred (except for contexts related to cannibalism).

These considerations support the usefulness of including information about selectional restrictions within a strategy of linking. My linking strategy will make use of such information via semantic filters (cf. section 6.5).

Jackendoff's treatment of selectional restrictions (and incorporations) as conceptual elaborations of thematic roles exhibits substantial commonalities and at the same time substantial differences with their treatment in EuroWordNet. In the next section, I will discuss the rationale behind the encoding of thematic roles in EWN. In general, thematic role relations can represent both selectional constraints and incorporations (cf. section 5.4). Hence, the uniform modelling of selectional restrictions and incorporations as information about a verb's thematic roles is common in Jackendoff's semantics and EWN. However, there is a crucial representational difference between the two approaches, which has remarkable theoretical implications. Jackendoff represents thematic roles and selectional restrictions by the same means: conceptual (sub-)structures. This suggests the view that thematic roles and selectional restrictions are basically the same kind of linguistic phenomena, and that the difference between

[16] Of course, the variant that both roles are A-marked has to be encoded as well. However, for the argument here, it is crucial that the two possibilities of marking only the Actor or only the Patient exist.

the two notions is just their degree of specificity (selectional restrictions are specifications of thematic roles). In contrast, EuroWordNet represents the two notions by different means: thematic roles are encoded as relations, whereas selectional constraints are encoded as noun concepts. This implies that roles and selectional restrictions are completely different in nature. Roles are a kind of *relational* entities, while selectional restrictions are a kind of *conceptual* entities. Of course, both phenomena are closely related; thematic roles are relations between predicates and their arguments, while selectional constraints represent conceptual restrictions holding for these arguments. However, in the wordnet paradigm, it would be inadequate to say that selectional restrictions are mere refinements of thematic roles, as in Jackendoff's approach.

2.3 Relevance for the Task of this Thesis

Given the different theories of thematic roles and selectional restrictions described in the previous sections, the crucial question is to what extent these theories are relevant for the task of learning thematic role relations for wordnets. I have tried to comment on certain aspects of the respective theories regarding this issue. In short, a concluding answer to this question is: While large parts of the detailed analyses of thematic roles and selectional restrictions (essentially idiosyncratic parts of the respective theories) are not applicable, other, more general insights (which are widely uncontroversial for all quoted theories) can be fruitfully employed for this task.

2.3.1 Thematic Roles and EuroWordNet

Obviously, none of the sketched analyses of the nature of thematic roles can immediately be applied here, because the information which would be required for that is not avaliable. Corpora (at least corpora of the size necessary for statistical learning) neither provide information about the deep structure or the situative saliency conditions underlying an utterance, nor the conceptual structure of a sentence and its parts, nor proto-role entailments which are valid for verb complements. (Euro)WordNet does not encode these (or related) kinds of information either. Here, thematic roles are encoded as relations between verb and noun concepts, labelled with an appropriate role type. This representation abstracts from the different views of the nature of roles. Thus, in a sense, EuroWordNet can be regarded as theory-open in this respect; it is open to different detailed analyses.

The major drawback of this abstract modelling is that it does not determine or facilitate the decision on the set of role types. This set has to be stipulated independently. Since the role inventories are different in the respective theories (although there is a considerable overlap), this stipulation is not straightforward. It is beyond the scope of this thesis to contribute to the discussion about an appropriate role inventory. I decided to adopt the role types which are defined in EuroWordNet. As noted in section 1.3, this decision is mainly motivated by a practical reason: To evaluate different learning approaches, I employ the role relations encoded in EuroWordNet as a gold standard, i.e. I compare them with the learned role relations (cf. section 5.4). Therefore, the types of the acquired roles have to correspond to the role types which are present in the gold standard.

As mentioned in section 1.2.3, the thematic role links from verb concepts to noun concepts in EWN are labelled with INVOLVED and the respective role type. The following list (cf. (Alonge 1996, p. 31–36)) enumerates the different INVOLVED relation types and test sentence templates which characterise the respective roles (X refers to a noun, Y to a verb in gerundive or infinitive form):

INVOLVED (A/An) X is the one/that who/which is involved in Y.

INVOLVED_AGENT (A/An) X is the one/that who/which does the Y.

INVOLVED_PATIENT (A/An) X is the one/that who/which undergoes the Y.

INVOLVED_INSTRUMENT (A/An) X is either i) the instrument that or ii) what is used to Y (with).

INVOLVED_LOCATION (A/An) X is the place where the Y happens.

INVOLVED_DIRECTION It is possible to Y from/to a place.

INVOLVED_SOURCE_DIRECTION (A/An/The) X is the place from which Ying happens / one Ys.

INVOLVED_TARGET_DIRECTION (A/An/The) X is the place to which Ying happens / one Ys.

(Alonge 1996) describes the rationale behind the choice of these role types as follows:

> The specific subrelations chosen, then, have not 'arbitrarily' been chosen, but are the most prominent ones to describe the different semantic references involved in the meaning of a word. Indeed, besides providing the possibility of encoding data on involved agents and patients, which are clearly relevant with respect to every class of verbs, these links both allow coding data on every kind of means used to perform an action (in fact 'instrument' is used here in a wide sense), and assume Gruber (1976) view according to which the semantics of motion and collocation can be seen as providing an interpretation for many semantic classes of verbs (besides the motion ones), e.g. verbs of *bringing*, *saying*, *giving*. (Alonge 1996, p. 31–32)

Thus, this role inventory is largely based on Gruber's and Jackendoff's perception of thematic roles. However, in some points, this inventory deviates significantly from that perception. First, there is a general role type INVOLVED (which was introduced presumably for pragmatic reasons, to provide a possibility to encode roles which differ from the other types). This resembles Fillmore's Objective case, which is also a kind of default case. Jackendoff explicitly rejects such a default role type, since in his theory, each role type has to correspond to a particular type of argument within conceptual structure. Second, one basic role type in Gruber's and Jackendoff's inventory is missing: Theme. Obviously, Themes are intended to be captured by INVOLVED_AGENT, INVOLVED_PATIENT, or INVOLVED_INSTRUMENT, respectively. This type does not play a role in Fillmore's account. Dowty (cf. (Dowty 1991, p. 555)) expresses doubts whether it is well-defined. Third, Jackendoff's distinction between thematic tier and action tier is not adopted for EWN.[17]

Thus, the role set in EuroWordNet exhibits properties of all three mentioned theories. And of course, it reflects the overlap between the role inventories of these theories, since it includes common role types such as Agent, Patient, and Instrument.

[17] Actually, this distinction could be represented by two different role relations (e.g. INVOLVED_TARGET_DIRECTION and INVOLVED_PATIENT) between a verb and a noun concept. However, this possibility is not pursued systematically.

Although the comprehensive theories of thematic roles cannot be employed for the task of this thesis as a whole, these theories provide very helpful insights concerning an important subtask: linking. It is worth noting that these insights are either very close to each other or complementary, in any case largely uncontroversial among the different theories. They relate to different kinds of information which provide clues for mapping syntactic complements to thematic roles:

- *Lexical information*
  Jackendoff's lexical entries make explicit that there is a correspondence between certain prepositions and certain roles of location and direction. For example, "from" indicates Source, while "to" or "into" indicate Goal.
- *Syntactic information*
  Fillmore states some isolated rules for mapping grammatical relations to deep cases, e.g. if an Agent is present, then it is realised as subject. Jackendoff and Dowty provide thematic hierarchies which integrate a bundle of such rules.
- *Semantic information*
  As Fillmore points out, different roles have different semantic restrictions (or, as I will argue in the next section, semantic preferences). E.g. Agents are likely to be animate, Instruments inanimate. Analogously, Dowty's Proto-Agent entailments of volitional involvement and sentience imply animacy.

Given a parsed corpus and WordNet, all three kinds of information are available (or at least deducible) for the individual syntactic complements to be linked roles during the learning process. For PP complements, the recognition of the preposition is straightforward. For NP complements, the grammatical relation is determined by the syntactic configuration of the sentence in question. Finally, the semantic concept corresponding to a complement can be acquired more or less reliably by word sense disambiguation techniques. This has to be done anyway in order to connect the word forms in the corpus to appropriate WordNet concepts. We will encounter several ways to achieve this. The approach which I applied is described in section 5.2.

Thus, it seems reasonable to devise a strategy for linking which combines these different kinds of information in a useful way. I will address this issue in chapter 6.

2.3.2 Selectional Preferences

In section 1.1.2, I mentioned the general idea to model selectional restrictions as preferences rather than as hard constraints. This idea, which I adopt for my work, was first put forward by Wilks (cf. (Wilks 1986)). This section discusses it from a linguistic point of view.

As explained in section 1.3, one of the core tasks of this thesis is the acquisition of selectional preferences by means of WordNet and the statistical analysis of corpora. The notion of *selectional preferences* was introduced by Resnik (cf. (Resnik 1993)) and generally adopted by other researchers who have been working on this task (cf. chapter 3). In a sense, this notion conflicts with the concept of selectional restrictions in the linguistic literature cited in section 2.2. In transformational grammar as well as in Jackendoff's conceptual semantics, selectional restrictions are *categorical*: either an argument conforms to the selectional restrictions of its predicate, or it violates them. In contrast, in the corpus-based, NLP-related research initiated by Resnik, selectional preferences are *graded*: They are

represented by numerical preference values which are assigned to WordNet concepts. A preference value quantifies how strong the corresponding WordNet concept is preferred (or dispreferred) by the predicate in question.

Now the question could arise whether this shift from a categorical to a graded view of selection should just be regarded as a secondary fall-out of the statistical acquisition approach (the preference values are determined by probabilities, so they are real-valued), or whether this graded view may be justified by reasons related to the phenomenon of selection itself. Like (Resnik 1993, p. 45–52), I will argue for the latter.[18]

As mentioned, (Chomsky 1965) discusses examples where the violation of selectional restrictions are not unnatural ((2.67)–(2.68) on page 36) if the respective constituent is embedded in a certain sentential context (typically of report or belief). In these cases, the embedding context constitutes a kind of "meta level" which either abstracts from the semantics of the embedded anomalous constituent, or refers to its semantic anomaly itself. From a corpus linguistic point of view, such constructions seem rather artificial. They do not play a major role in human communication. There are much more prevalent circumstances which involve the violation of selectional restrictions, or, to put it another way, suggest a more flexible approach to classify instances of selection than choosing between the two possibilities 'satisfying' and 'violating'. In the following, I will discuss two of these circumstances: metaphor and specific contexts.

Metaphorical expressions typically imply the violation of selectional restrictions. For example, "drink" selects a human being or an animal as its Agent. However, in the sentence

(2.98) Fast cars drink petrol.

found in the British National Corpus (text CMM, sentence 890), the Agent is a car. In the traditional analysis, this is regarded as a violation of selectional restrictions. However, there is a straightforward metaphorical interpretation of the sentence, namely 'Fast cars consume (a lot of) petrol'. This is compatible with the categorical model of selectional restrictions. Chomsky himself points out that violations of selectional restrictions can often be interpreted metaphorically or within an appropriate context (Chomsky 1965, p. 149). However, the categorical model has no means to distinguish between instances like (2.98) and cases which apparently lack a natural interpretation, like

(2.99) Modern sunglasses drink petrol.

If, however, we regard selection as preferential, then we can describe this difference by assigning different values of (dis)preference to the different concepts. The selectional preferences of "drink" for its Agent can be modelled by values which indicate high preference for the concepts <human being> and <animal>, low preference (or weak dispreference) for the concept <car>, and strong dispreference for concepts like <sunglasses>.

The crucial point is that if a concept is dispreferred, then this does not mean that it is completely impossible. As McCawley emphasises, semantically anomalous constituents are not meaningless (cf. examples (2.84)–(2.86) on page 41). In other words, an expression containing a violation of selectional restrictions nonetheless has a (maybe very fragmentary) meaning. This meaning may or may

[18] However, I will emphasise different aspects than Resnik's discussion does.

not suggest a metaphoric interpretation.[19] However, for basically any semantically anomalous expression one can find (or construct) a more or less specific context in which it seems appropriate. To illustrate this, Resnik provides a context for the phrase "buxom father", the introduction of a guest in a talk show:

(2.100) "Now introducing John Smith, looking lovely after his breast-augmentation surgery. This buxom father of two makes his living in Las Vegas as a female impersonator..." (Resnik 1993, p. 46)

This context is exceptional, but still possible in the real world. In a fictional setting, one can imagine things like sunglasses whose owner always wears them during motor-biking, and which finally start drinking petrol in order to impress his noisy and arrogant Harley-Davidson. It is even not a too difficult exercise to construct a context fitting to sentences like "Colorless green ideas sleep furiously.", as illustrated, for example, by Yuen Ren Chao's well-known "Story of my Friend, Whose Colorless Green Ideas Sleep Furiously".[20]

For these reasons, it is more appropriate to model selectional phenomena as preferences rather than strict constraints. This makes it possible to regard semantically anomalous constructions as (more or less strongly) dispreferred, instead of rejecting them completely. In this way, metaphoric use and different contexts can be handled with more flexibility.

It should be emphasised that these issues are anything but empirically marginal. The context-dependence of semantic interpretation (and hence, selectional preferences) is pervasive. We have encountered already several examples which are relevant in this respect. In sentence (2.83) on page 40, it highly depends on the context in which the sentence is uttered whether "bachelor" is interpreted as 'who has the first or lowest academic degree' or 'who has never married'. In terms of selectional preferences, the predicate "be a bachelor" in the sense 'be a male single' generally prefers male human beings as its argument. However, given a context in which the marital status of people is highlighted and their sex is secondary (e.g. when talking about the advantages and disadvantages of being married), then this predicate essentially prefers male or female human beings (although males still may be stronger preferred than females), so that the 'single' sense of "bachelor" is not ruled out by the argument "aunt". Moreover, in such contexts, this sense is more salient than the 'academic degree' sense.

Another example where a specific context influences selectional preferences is (2.96) on page 44, Generally, "cook" prefers a human as its Agent and food (originating from animals or plants) as Patient. Therefore, "mother" is clearly interpreted as Agent. But in a context related to cannibalism (which is, unfortunately, not as exotic as it might seem, since terrible crimes of that kind do happen in our society), a "thematic ambiguity" arises: humans are preferred for the Agent *and* the Patient of "cook"; thus, "mother" could be Agent or Patient.

Obviously, selectional preferences acquired by statistical approaches reflect the quantities in the examined corpus. The preference value of a certain noun concept w.r.t. an argument of a certain verb

[19] I will not go much into details about which factors constitute metaphor. This would go far beyond the scope of this chapter. I refer the interested reader to (Lakoff & Johnson 1980) for a theoretical discussion of metaphor and (Fass, Martin & Hinkelman 1992) for AI approaches to formalise non-literal language.

[20] In the context established by this "story", metaphoric interpretation necessarily plays a major role, which illustrates that the issues of metaphoric uses and specific contexts relate to each other.

depends on the relative co-occurrence frequency of these two. Evidently, this frequency is significantly influenced by the contexts in which the verb occurs. And this, in turn, largely depends on the composition of the examined corpus. Thus, it is possible to acquire selectional preferences specific to a particular domain just by restricting the data to texts which originate from that domain. (For example, investigating a corpus of fairy tales might yield animals, plants, and mirrors as preferred Agents of "say".) If one wants to learn selectional preferences to supplement a "general-purpose resource" like WordNet, a balanced corpus of reasonable size (like the British National Corpus) is the appropriate choice.

The presence of metaphoric language is not less significant. (Lakoff & Johnson 1980) point out the ubiquity of metaphor in language and thinking:

> ...metaphor is pervasive in everyday life, not just in language but in thought and action. Our ordinary conceptual system, in terms of which we both think and act, is fundamentally metaphorical in nature." (Lakoff & Johnson 1980, p. 3)

There is a priori no correlation between the degree of preference on the one hand and literal vs. non-literal language on the other hand. The discussion of example (2.98) above could be misleading, since it may suggest that metaphoric use tends to coincide with a low preference value. This would imply that non-literal language is significantly less frequent than literal language. However, this is not necessarily the case, and, in the light of Lakoff's and Johnson's statement, not even what one would expect.

(Lakoff & Johnson 1980) identify conventionalised metaphors on the conceptual level. These *conceptual metaphors* have more or less common expressions as linguistic instances. For example, TIME IS MONEY is a conceptual metaphor underlying utterances like

(2.101) You're wasting my time.

(2.102) How do you spend your time these days?

(2.103) The flat tire cost me an hour.

(2.104) I've invested a lot of time in her.

(2.105) You need to budget your time.

It is a purely empirical question whether such expressions occur with a significant frequency to be reflected by correspondingly high preference values. For example, it might turn out that the concept <time> receives a rather high preference value for "invest", but not for "budget". Furthermore, not all metaphoric expressions are instances of a systematic conceptual metaphor. There are "isolated" conventionalised metaphoric expressions as well as novel, "inventive" metaphors. The representation of selectional preferences as numerical values attached to concepts is not able to indicate such differences, nor to distinguish between literal and figurative language at all.

But, indeed, this representation does not aim at such a level of differentiation. Nor does it *explain* why "drink" prefers <car> much more than e.g. <sunglasses> (i.e. that petrol is liquid, a car is a thing which consumes petrol, and drinking is a kind of consuming), or which contextual conditions favour the interpretation of "bachelor" as 'male or female single'. In general, such explanations would

include an illimitable set of inferences on world knowledge. A statistical representation of selectional preferences abstracts from such facts and inferences[21] and, instead, provides a level of information which allows interesting insights into *quantitative aspects* of selection. A complex system comprising a set of facts and inferences would not supply these insights.

This point is put forward by (Abney 1996*b*):

> If we have a complex deterministic system, and if we have access to the initial conditions in complete detail, so that we can compute the state of the system unerringly at every point in time, a simpler stochastic description may still be more insightful. To use a dirty word, some properties of the system are genuinely *emergent*, and a stochastic account is not just an approximation, it provides more insight than identifying every deterministic factor. Or to use a different dirty word, it is a *reductionist* error to reject a successful stochastic account and insist that only a more complex, lower-level, deterministic model advances scientific understanding. (Abney 1996*b*, p. 19)

Abney argues that statistic data are not only an issue of performance, and hence useful for NLP tasks, but also relevant for theoretical linguistics. He sketches how statistical information could be fruitfully employed in different areas of linguistics, and concludes:

> Statistical methods—by which I mean primarily weighted grammars and distributional induction methods—are clearly relevant to language acquisition, language change, language variation, language generation, and language comprehension. Understanding language in this broad sense is the ultimate goal of linguistics. (Abney 1996*b*, p. 21)

Meanwhile, statistic (or at least quantitative) methods are established in linguistic research, e.g. in employing corpus frequencies to compare different constructions, or for setting up and evaluating psycholinguistic experiments in which the informants are asked for (graded!) judgements e.g. about the grammaticality of certain sentences. Thus, a graded model of selection, which has emerged from the renaissance of statistical methods in computational linguistics, is as well conformable with a recent tendency in theoretical linguistics.

To conclude the discussion, I would like to point out that the categorical model and the graded model of selectional restrictions do not mutually exclude each other. Referring to (Drange 1966), (Resnik 1993, p. 47) brings up the possibility that selection is a matter of category *and* a matter of degree. In fact, there is evidence for this view.

Recall Jackendoff's discussion of the violation of selectional restrictions in (2.92) and (2.93) on page 43. Jackendoff notes that "drink powder" contains a clash between the conceptual features SOLID, which is part of the conceptual structure of "powder", and LIQUID, which "drink" selects for its Theme. The major category of "powder" and the Theme of "drink" agree, it is Thing in both cases. However, in "drink sincerity", there is already a clash of major category, since "sincerity" has the category Property. Although Jackendoff does not explicitly note it, different types of conceptual clash

[21] Note that a categorical model of selection involves this abstraction as well. Following (Johnson-Laird 1983), Resnik characterises selectional constraints as "inferences conventionalised because of their frequency and predictability" (Resnik 1993, p. 48). This is in a sense analogous to Jackendoff's view of selectional restrictions as "convenient name" rather than a peculiar theoretical concept.

(with different severity) suggest that there are different degrees of violation of selectional restrictions (though here the scale of violation would not be real-valued but finite). Nevertheless, Jackendoff's account is categorical: there is a clear distinction between the violation of and the conformance to selectional restrictions.

Although Jackendoff's conceptual framework to a certain extent accounts for degrees of dispreference, it does not capture degrees of preference. However, it is not difficult to find obvious examples of a verb preferring different noun concepts with different strengths. Consider

(2.106) The pupil read the article.

"article" may refer to a word (a determiner) or to a text. "read" prefers both senses (you can read a word as well as a text), but the 'text' sense is much more typical, and thus preferred much stronger than the 'word' sense.[22]

Some of the approaches introduced in chapter 3 calculate the preference values in a way that allows a clear distinction between preferred and dispreferred concepts. For example, in some approaches a positive value indicates preference, a negative value dispreference; the absolute value indicates the degree of preference or dispreference, respectively. In this case, one could imagine that to model the selectional preferences for the Patient of <read>, the concept <text> would get the preference value 12.2, <word> the value 1.9, <picture> -0.9, and <hammer> -6.7. These approaches integrate the categorical and the graded view of selection. In the example just mentioned, the preference values would imply that <read> prefers both <text> and <word>—but the former concept is more preferred—and disprefers both <picture> and <hammer>—but the latter is more dispreferred. Other approaches do not draw a clear line between preference and dispreference. Since my goal is to integrate the acquired selectional preferences into WordNet, and since the WordNet data model does not include quantitative information attached to relations,[23] I have to employ a method which draws a clear line between preference and dispreference, since verb concepts should only receive thematic role relations to noun concepts which they actually prefer. In other words, since WordNet encodes categorical rather than graded information, the approach I employ has to capture the categorical aspect of selection. Nevertheless, the WordNet relations acquired in this way encode preferences rather than restrictions.

[22]These are introspective judgements to make the point clear. The statistical approaches advocated here just aim at providing an empirical basis for such judgements.

[23]Of course, this does not mean that the preference values ultimately have to be dropped completely. It is no problem to provide the values for the preferred concepts (as well as information concerning dispreferred concepts) as extra information which is not integrated in WordNet.

Chapter 3

Acquiring Selectional Preferences from Corpora: Existing Approaches

In chapter 1.3, I have formulated the task of this thesis as developing an approach for extending WordNet with thematic role relations such as

(3.1) <drink> INVOLVED_AGENT <human>

(3.2) <drink> INVOLVED_AGENT <animal>

(3.3) <drink> INVOLVED_PATIENT <liquid>

I have explained that due to the tight relationship between thematic roles and selectional restrictions, this task intrinsically requires to employ a method of acquiring selectional preferences from corpora. In recent years, several approaches have been proposed in the literature for learning selectional preferences by means of statistical corpus analysis. In this chapter, I will review such approaches. In particular, I will focus on properties and design decisions which turn out to be crucial for the development of the approach that I propose in this thesis. For this reason, I first outline some criteria for the suitability of a method for learning thematic role relations (section 3.1). After that, I describe the kind of training data which in principle is common to all approaches reviewed in this chapter, as well as the method I develop in this thesis (section 3.2). Then I turn to the presentation of individual approaches. In particular, I explain one method which does not employ a lexical resource like WordNet as background knowledge (section 3.3) and a number of approaches which are based on WordNet (section 3.4). I will examine to what extent these learning methods meet the suitability criteria to be outlined. In section 3.4.7, I will provide a summary of the WordNet-based approaches and argue that the approach introduced in section 3.4.3 fits these criteria best. This approach forms the starting point for my development of a thematic role acquisition method.

I will assume that the reader is familiar with the principles of probability theory. Furthermore, I will not provide a general introduction to common techniques in statistical NLP such as Hidden Markov Models (Rabiner 1989) or the EM algorithm (Dempster, Laird & Rubin 1977). These techniques are presented in various introductions to the field (e.g. (Charniak 1993), (Mitchell 1997), or (Manning & Schütze 1999)).

A general notional remark: In the formulae presented in this and the following chapters, it is necessary to distinguish between word forms[1] on the one hand and word concepts on the other hand. I will make this distinction explicit by using different variables. The variables v and n, respectively, will be employed to refer to verb or noun forms. In contrast, $vcpt$ and $ncpt$, respectively, will usually be employed to refer to verb or noun concepts (i.e. synsets), except for one special case: Recall that wordnet synsets represent at the same time word senses and abstract concepts (cf. section 1.2.1). In some contexts, one of these two representational functions is dominant. In particular, it will be important to explicitly mark situations where synsets primarily represent word senses rather than semantic abstractions. In those cases, the synsets in question will be referred to by the variables $vsns$ and $nsns$, respectively.

3.1 Suitability Criteria

In order to be suitable for integration in wordnets as thematic role relations, the preferences acquired by statistical methods have to be in a certain form. In particular, they have to

1. be expressed as relations between verb concepts and noun concepts in WordNet
2. represent the appropriate level of generalisation
3. allow the distinction between preferred and dispreferred concepts

Regarding the first criterion, it is crucial to note that some statistical methods for acquiring selectional preferences proposed in the literature are not connected to WordNet at all, such as clustering approaches as the one described in section 3.3. These clustering methods model selectional associations between verb and noun forms (i.e. lemmas) rather than verb and noun concepts. For this reason, they are not immediately applicable for learning EWN-like thematic role relations. (Nevertheless, as we will see in chapter 5 and 6, such approaches turn out to be very useful within several preprocessing steps.) The methods described in section 3.4 acquire preferred (and dispreferred) noun concepts for a given verb form, i.e. they represent selectional preferences as relations between verb forms and noun concepts. Only one of these approaches explicitly aims at acquiring selectional preferences of verb concepts as well. However, all approaches could be extended accordingly: As the nouns in the training corpora are not semantically disambiguated, certain techniques are employed for mapping the noun forms in the data to corresponding WordNet noun concepts. Analogous techniques could be applied to the verb forms. Then, selectional preferences can be acquired for verb concepts using the same learning algorithm as for verb forms.

The second criterion has been mentioned already in chapter 1. It states that the noun concepts which are acquired as preferences of a certain verb concept have to be at a level of abstraction that is *empirically adequate* (i.e. captures all and only the preferred concepts) on the one hand and *as compact as possible* on the other hand. For example, it is inappropriate to introduce a PATIENT relation from <eat> to a concept which is so general that it also subsumes dispreferred concepts (e.g. <entity>).

[1] For terminological clarity, I would like to point out that the term *word form* in this thesis is used in opposition to the notions of word sense and semantic concept, i.e. it refers to an occurrence of a word in the training corpus, which is not semantically disambiguated. It is *not* used in a morphological sense, i.e. it does not refer to a full form as opposed to a lemma. In the training data used for all approaches described in this thesis, verbs and nouns are lemmatised.

Likewise, it is not desirable to establish relations to all the many food concepts in the wordnet (<potato>, <milk_chocolate>, <mock_turtle_soup>,...) because such a representation would be highly redundant and inefficient. Moreover, it would fail to express any generalisation. The demand to avoid acquiring concepts which are too specific is crucial not only w.r.t. storage and processing economy, but also for functional reasons: Firstly, expressing appropriate generalisations is important for certain NLP applications like semantic inferencing. Secondly, a concise encoding of selectional preferences enhances readability for human users. In the ideal case, a concept can be found which just subsumes all the preferred and none of the dispreferred concepts (such as <food>).

It is important to point out that these general considerations should be regarded relative to the needs of a particular application and/or approach. In different situations, different levels of abstraction are adequate. Two examples might illustrate this point: On the one hand, there are parsing systems (e.g. (Huyck 2000)) which make use of selectional restrictions modelled by a small set of very abstract semantic classes (like *human*) to resolve structural ambiguities. On the other hand, there are lexical resources (e.g. FrameNet, cf. (Fillmore et al. 2001)) where selectional information is represented by nouns which *typically* occur at a certain frame slot of some verb. Thus, these resources model selectional preferences at a rather low level of generalisation. This is appropriate for particular NLP applications, e.g. in a text understanding system tuned to a particular domain (cf. (Fillmore & Baker 2001)). A reusable approach for acquiring selectional preferences should be adjustable to such application-specific *a priori* design principles regarding the desired degree of generalisation, and, *within this degree*, behave as sketched in the previous paragraph.

The third criterion means that the acquired preference values should model not only the graded aspect of selectional preferences, but also their categorical aspect. At the end of section 2.3.2, I discussed this issue. Statistical methods as described in section 3.4 model the preference or dispreference of a noun concept w.r.t. a certain verb by a numeric preference value, i.e. a real number assigned to that concept. Criterion 3 states that a numerical limit should exist which splits the range of possible preference values into a subrange of values indicating preference and a subrange of values indicating dispreference (e.g. values above 0 indicate preference, values below 0 dispreference). Approaches which provide such a limit in a principled way are favourable. For the task of acquiring thematic role relations, it is necessary to make the categorical distinction between preference and dispreference, since only relations from verb concepts to *preferred* noun concepts should be integrated in a wordnet.

I will discuss to which extent the different WordNet-based approaches meet the criteria 2 and 3 in section 3.4.

3.2 Training Data

All approaches described in this chapter acquire selectional preferences for syntactic complements of verbs, i.e. there is no attempt to link these complements to thematic roles. The required training data are obtained from a parsed corpus. These training data comprise a collection of verb–complement co-occurrence triples (v, r, n) where v is a verb, r is a syntactic relation, and n is the head noun of the NP complement which is related to v via r. Each triple corresponds to an instance of a syntactic (sub-)structure in the corpus. The syntactic relation r can be subject, (direct) object, indirect object, or a certain preposition. In the latter case, r is the preposition of a PP complement of v and n is the head

(eat, subject, animal)
(eat, subject, child)
(eat, object, banana)
(eat, object, lunch)
(eat, object, potato)
(come, from, newspaper)
(come, from, room)
(come, to, end)
(come, to, office)

Figure 3.1: Examples of verb–complement co-occurrence triples

of the NP inside the PP. Figure 3.1 shows some examples of such (v, r, n) triples.[2] These examples are extracted from the British National Corpus (BNC) (cf. (Burnard 1995)), which is employed by some of the approaches described here.

From this set of (v, r, n) triples, the approaches described here are able to learn selectional preferences of each verb v for each syntactic relation r it subcategorises for. Note that the structure of these data implies two simplifications with respect to the linguistic findings about selectional restrictions discussed in section 2.2. First, it does not account for dependencies between the preferences of a verb for its different arguments. Chomsky notes that such dependencies exist (cf. example (2.61) on page 35). For example, if the object of "eat" is "mouse", then the subject is much more likely to be "cat" or "buzzard" than if the object is "sandwich". Such dependencies are neglected in order to avoid a sparse data problem: A statistical model which captures this kind of dependencies would be much more complex and the number of required parameters would be multiplied. In addition to marginal probabilities for single complements of a verb, e.g. $p(subj = cat|eat)$ and $p(obj = mouse|eat)$—these marginal distributions are estimated by the approaches described in this chapter—one would need an estimation of joint probability distributions over different arguments of a verb, like $p(subj = cat, obj = mouse|eat)$, to obtain the corresponding conditional probabilities (e.g. $p(subj = cat|obj = mouse, eat)$). It is obvious that much more data are required to reliably estimate the co-occurrence probabilities of three or more words (a verb and two or more complement nouns) than to estimate co-occurrence probabilities of just two words (a verb and one complement noun). Li & Abe (1996) examined a model which captures pairwise dependencies between verb complements. They report that training this model on the Penn Treebank (Marcus, Santorini & Marcinkiewicz 1993) did not provide evidence for any such dependencies.[3] This supports the claim that assuming independence between the complements of a verb is quite appropriate in practice. However, Li and Abe admit that the size of the training corpus could have been insufficient to detect dependencies. The Penn Treebank contains about 3 million words. In the middle of the 1990s, this was the usual magnitude for a parsed corpus. Nowadays, corpora with 100 million words (like the BNC) or more are available as well as robust parsing techniques to process them. It would be worth training Li & Abe's (1996) model on corpora of that size. I decided not to pursue this option, because

[2]Each triple represents a token, not a type of co-occurrence; thus, if e.g. "potato" occurs several times as the object of "eat" in the corpus, then the triple *(eat, object, potato)* occurs in the training data with the corresponding frequency.

[3]More precisely, they did find dependencies when employing a model that only takes the presence or absence of syntactic complement types into account, but not for models which comprise distributions over complement nouns (or corresponding WordNet concepts, respectively).

it would not serve the eventual task of learning thematic role relations, since in (Euro)WordNet it is not possible to encode dependencies between relations. Nevertheless, it would be interesting to find out if corpus size matters here.

The second simplification implied by the structure of the training data is that it does not capture the complete complement constituents, but only their heads. As illustrated by example (2.80) on page 40, the semantics of the entire complement is subject to a verb's selection. However, the attempt to take into account the complete complement phrase would involve two intractable problems: the appropriate semantic representation of a phrase and data sparseness. Approaches which represent selectional preferences in terms of WordNet concepts (section 3.4) have to map the word forms in the corpus to corresponding WordNet concepts. In this chapter and chapter 5, we will encounter several strategies to achieve such a mapping for simple noun forms. However, since the synsets in WordNet are intended to represent lexical items, concepts which are equivalent to the meaning of a complex phrase are not avaliable in general. This problem does not apply for clustering approaches which are not based on a lexical resource (cf. section 3.3). But for both kinds of approaches, the sparse data problem is striking: Theoretically, there is an infinite number of possible complement noun phrases, which means that for a reliable estimation of their probabilities, an unlimited number of data would be required. Even if one introduces an upper limit for the number of words in a phrase to be taken into account, the number of parameters to estimate grows exponentially with this upper limit. The decision to only take the head into account just means to set this upper limit to 1. Furthermore, it is reasonable to assume that in general the heads carry the main information of the complement phrases and thus provide a suitable base for acquiring a verb's selectional preferences.

3.3 Approaches without Employing Background Knowledge

There have been several proposals in the literature to characterise selectional associations between verbs and their complement nouns by clustering verb–noun pairs. In such a cluster model, the selectional preferences of a verb are indicated by the nouns which co-occur with that verb in the same cluster(s). Well-known approaches which implemented this idea include the work of Pereira, Tishby & Lee (1993) and Rooth et al. (cf. (Rooth 1998), (Rooth, Riezler, Prescher, Carroll & Beil 1998)). In the following, I will describe the latter approach, which the authors call *latent semantic clustering*. As I will point out below, this approach turns out to be useful for several parts of my work.

The probability model induced by Rooth et al. aims at clustering verb–noun pairs according to their selectional pattern. The idea behind the formation of such clusters is based on the assumption that selectional patterns constitute groups of verbs and groups of nouns so that the members of a certain verb group tend to select the members of a certain noun group. For example, verbs like "resolve", "face", "address", or "fight" select nouns like "problem", "pressure", "damage", or "challenge" as their object. Such a pattern corresponds to a cluster, a so-called *latent semantic class (LSC)*, that comprises (among others) these verbs and nouns. As indicated by the term *latent*, a latent semantic class is a soft cluster, i.e. the membership of a verb–noun pair (v, n) in a class c is graded. I will illustrate the rationale behind this approach by an example.

Figure 3.2 shows an (artificially constructed) latent semantic class. This class, labelled c_{ex}, represents verb–object pairs, e.g. $(face, problem)$, $(face, pressure)$, $(resolve, problem)$, $(resolve, pressure)$, etc. As stated above, the LSC model is a soft clustering method. Thus, verb–noun pairs (v, n) are members of a class c to a certain degree. This degree is quantified by the con-

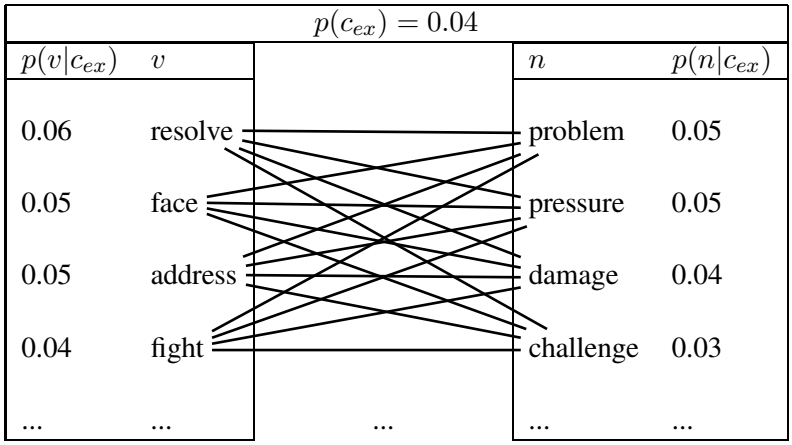


Figure 3.2: Example of a latent semantic class

ditional probabilities $p(v, n|c)$. The basic structural principle of the LSC approach is that these joint probabilities are not immediately represented as parameters of the model, but are determined by the marginal probabilities $p(v|c)$ and $p(n|c)$. These probabilities are represented in the model. Figure 3.2 displays the verbs with their probabilities on the left-hand side of the diagram (e.g. $p(resolve|c_{ex}) = 0.06$) and the nouns with their probabilities on the right-hand side (e.g. $p(problem|c_{ex}) = 0.05$). From these marginal probabilities, the joint probabilities for verb–noun pairs are determined as follows:

$$p(v, n|c) = p(v|c) \times p(n|c) \tag{3.4}$$

For example, for the pair $(resolve, problem)$ this equation yields the following probability of membership in class c_{ex}: $p(resolve, problem|c_{ex}) = p(resolve|c_{ex}) \times p(problem|c_{ex}) = 0.06 \times 0.05 = 0.003$. Technically, equation (3.4) states that $p(v|c)$ and $p(n|c)$ are independent. This independence assumption is the fundamental generalisation that an LSC model represents. Intuitively, a cluster gathers a group of similar verbs and a group of similar nouns so that these verbs and nouns tend to co-occur with each other. This means that if a verb v and a noun n co-occur in the data more often than by chance, then they are likely to be a member of the same class (more precisely: to receive a high membership probability in the same class). But additionally, if a verb v' is similar to v regarding the nouns with which it co-occurs, but does *not* co-occur at all with the specific noun n in the data, then v' is likely to be in the same cluster as v, and the pair (v', n) is predicted with a certain probability. This holds analogously for nouns. For example, it could be that "challenge" never occurs as object of "fight" in the data. However, these words are grouped in class c_{ex} together with the other verbs and nouns displayed in the figure. The LSC model predicts the membership of the pair $(fight, challenge)$ in the class with the probability $0.04 \times 0.03 = 0.0012$. Hence, the LSC approach is helpful to overcome the sparse-data problem w.r.t. verb–noun co-occurrences.

In addition to the membership probabilities for verbs and nouns, each class has a certain probability $p(c)$ in the LSC model. (In the example in figure 3.2, this probability is $p(c_{ex}) = 0.04$). With these probabilities, one can formulate an equation that estimates the overall probability of a verb–noun pair (v, n):

$$p(v,n) = \sum_c p(c,v,n) = \sum_c p(c) \times p(v|c) \times p(n|c) \tag{3.5}$$

Now I turn to the method of estimating an LSC model. The probabilities $p(c)$, $p(v|c)$, and $p(n|c)$ for each c, v, and n are the parameters of the LSC model. They are learned by maximum likelihood estimation from incomplete data via the EM algorithm. The observable (incomplete) data consist of verb–noun co-occurrence pairs $(v.n)$ which are retrieved from the training data described in section 3.2 (e.g. by extracting all verb–noun pairs connected by a particular grammatical relation). The corresponding unobservable (complete) data are (c,v,n) triples, i.e. verb–noun pairs together with the respective classes which they are instances of. The number of classes has to be arbitrarily fixed before clustering. The clustering procedure works as follows: In an initialisation step, the parameters are randomly chosen. After that, a certain number of training iterations (this number also has to be arbitrarily fixed) is performed which re-estimate the parameters. Each iteration consists of the E-step (estimation) and the M-step (maximisation). The E-step calculates the expected frequencies of (c,v,n) triples (the expectation values $E(c,v,n)$) by employing the corresponding probabilities $p(c,v,n)$ and $p(v,n)$ (which can be estimated from the parameters according to equation (3.5)) and the frequencies[4] of the respective pairs (v,n) in the observed data:

$$E(c,v,n) = freq(v,n)\frac{p(c,v,n)}{p(v,n)} \tag{3.6}$$

The related expectation values $E(c,v)$, $E(c,n)$, and $E(c)$ are computed straightforwardly:

$$E(c,v) = \sum_n E(c,v,n) \tag{3.7}$$

$$E(c,n) = \sum_v E(c,v,n) \tag{3.8}$$

$$E(c) = \sum_{v,n} E(c,v,n) \tag{3.9}$$

In the M-step, these expectations are used to re-estimate the model parameters by maximum likelihood estimation (MLE):

$$p(v|c) = \frac{E(c,v)}{E(c)} \tag{3.10}$$

$$p(n|c) = \frac{E(c,n)}{E(c)} \tag{3.11}$$

$$p(c) = \frac{E(c)}{\sum_{v,n} freq(v,n)} \tag{3.12}$$

(For a formulation of this estimation procedure in terms of the general framework of maximum likelihood estimation from incomplete data via the EM algorithm presented in (Dempster et al. 1977), cf. (Rooth et al. 1998, p. 99–101).)

[4] In this work, I will refer to the frequency of an item x by $freq(x)$.

$p(c_{29}) = 0.052$				
v	$p(v\|c_{29})$		n	$p(n\|c_{29})$
allow	0.05595		company	0.07287
give	0.04996		people	0.05791
tell	0.04769		investor	0.04724
help	0.04348		government	0.02802
ask	0.03068		customer	0.02600
require	0.02369		employee	0.02555
say	0.02168		worker	0.02435
attract	0.02134		client	0.01987
force	0.02101		bank	0.01928
represent	0.02067		shareholder	0.01402
protect	0.01823		agency	0.01363
leave	0.01805		consumer	0.01271
urge	0.01768		group	0.01270
keep	0.01565		state	0.01254
include	0.01482		reporter	0.01234
encourage	0.01475		court	0.01153
...	...		...	...

Figure 3.3: A latent semantic class which represents verb–object relations

Rooth et al. show two different ways to obtain the (v, n) pairs which form the input of their learning approach from the training corpus. (Rooth 1998) pursues the straightforward possibility of only taking into account one grammatical relation at a time. i.e. to extract those verb–noun pairs which are connected via a particular grammatical relation. An LSC model derived from these pairs represents selectional association with respect to this relation. (Rooth 1998) presents such a model for the direct object. An example of a class in that model is shown in figure 3.3 (the verbs and nouns are sorted according to their probabilities in descending order). This class groups together (activities involving) commercial products.

A more sophisticated alternative is demonstrated in (Rooth et al. 1998). Here, all grammatical relations are taken into account at once to induce an LSC model. To achieve this, the information provided by the left part of a (v, n) tuple is extended: In addition to a verb form, this component contains the complete subcategorisation frame of the corpus sentence which corresponds to the tuple. Furthermore, that grammatical relation which connects the verb to the noun at the right part of the tuple is explicitly marked. Figure 3.4 shows an example of a cluster obtained from such kind of data (the respective marked grammatical relations are underlined).

Concerning the task of this thesis, the classes in the two figures illustrate interesting properties exhibited by LSCs. First, they tend to group together semantically similar words. The class in figure 3.3 groups verbs which denote giving orders or permissions, like “allow”, “tell”, “ask”, and “urge”, as well as nouns referring to organisations, like “company”, “government”, and “bank”, or to (in a sense dependent) people, like “customer”, “client”, “employee”, and “worker”. The class in figure 3.4 groups scalar change events. It contains similar verbs like “increase”, “rise”, “fall”, and “reduce”, as well as similar nouns, like “number”, “rate”, and “amount”, or “price” and “cost”. As can be seen, it is

$p(c_{17}) = 0.0265$			
v	$p(v\|c_{17})$	n	$p(n\|c_{17})$
increase (subj)	0.0437	number	0.0379
increase (subj obj)	0.0392	rate	0.0315
fall (subj)	0.0344	price	0.0313
pay (subj obj)	0.0337	cost	0.0249
reduce (subj obj)	0.0329	level	0.0164
rise (subj)	0.0257	amount	0.0143
exceed (subj obj)	0.0196	sale	0.0110
exceed (subj obj)	0.0177	value	0.0109
affect (subj obj)	0.0169	interest	0.0105
grow (subj)	0.0156	demand	0.0103
include (subj obj)	0.0134	chance	0.0099
reach (subj obj)	0.0129	standard	0.0091
decline (subj)	0.0120	share	0.0089
lose (subj obj)	0.0102	risk	0.0088
act (subj obj)	0.0099	profit	0.0082
improve (subj obj)	0.0099	pressure	0.0077
...	...	...	...

Figure 3.4: A latent semantic class which represents any kind of grammatical relations

generally *not* the case that the words in a cluster can be subsumed by *a single* semantic feature—the nouns in figure 3.3 require at least two, namely 'company' and 'person'. Nevertheless, the similar words which are grouped within one class provide evidence for mutual disambiguation. For example, in figure 3.3, nouns like "agency", "company", or "government" support the organisation sense of "bank", whereas the waterside sense does not have similar counterparts in that class. This suggests to employ latent semantic clustering for preprocessing the training data in order to (partially) disambiguate them and improve the mapping from word forms to WordNet concepts. As I will describe in section 5.2, I applied this technique for that purpose.

Figure 3.4 shows another interesting property: When all grammatical relations are taken into account, latent semantic clustering tends to group together corresponding syntactic realisations of a particular semantic argument of a verb. For example, "increase (subj)" and "increase (subj obj)" are grouped in the same class. Both syntactic configurations realise the Patient of "increase". Information about semantically corresponding syntactic patterns in which a verb and its complement occur is very useful for the task of linking. In chapter 6, I will describe how I employed this information for mapping syntactic complements to thematic roles. Thus, although the approach of Rooth et al. cannot be directly adopted for the task of learning thematic role relations (since it does not involve wordnets at all), it turns out to be very helpful for two important sub-tasks.

3.4 Approaches Employing WordNet

At the beginning of the 1990s, Philip Resnik initiated research on acquiring selectional preferences by statistical corpus analysis based on WordNet. Following his work (Resnik 1993), several approaches for that task have been proposed (e.g. (Ribas 1994), (Abe & Li 1996), (Li & Abe 1998), (Agirre & Martinez 2002), (Clark & Weir 2002), (Abney & Light 1999)). In this section, I will describe a variety of these approaches in order to point out several design decisions which are crucial with respect to the task of this thesis, in particular the suitability criteria mentioned in section 3.1.

3.4.1 Resnik

3.4.1.1 Mapping the Training Data to WordNet: the Word-to-Concept Approach

The statistical methods described here express selectional preferences in terms of probabilities of WordNet concepts, which are estimated on the basis of evidence provided by the training data. The preliminary step which is necessary to achieve this is to constitute a mapping from the word forms in the data to the WordNet concepts which correspond to these word forms. This mapping aims at estimating frequency counts of concepts, which in turn are employed to derive concept probabilities. I mentioned above that most of the work which is relevant here acquires selectional preferences for verb forms rather than verb concepts so that such a mapping is required for nouns only. However, it can be performed for verbs as well.

Since the nouns in the training data are not disambiguated, it is not possible to decide which sense a certain noun instance represents. Lacking this information, the mapping mechanism has to treat a noun as an indication of all concepts which represent one of its possible senses. For example, "bank" provides evidence (among others) for the concepts <bank#banking_company> and <bank#side>. In addition, a noun provides evidence for the (direct and indirect) hyperonyms of these concepts, since the semantics of the WordNet hierarchy implies that a concept subsumes all its subconcepts. Thus, "bank" also indicates the concepts <financial_institution>, <institution>, <organization>, <social_group>, and <group>, as well as <incline#side#slope>, <geological_formation>, <natural_object>, <physical_object>, and <entity>.

Resnik (1993) proposes a mapping approach which is based on these considerations. He divides the frequency count of a noun n equally among all concepts $concepts(n)$ for which the noun provides evidence, i.e. the concepts representing a sense of the noun and their hyperonyms.

Figure 3.5 illustrates how this approach works. Suppose that the word "someone" occurs 100 times in the training sample. There are four WordNet concepts that subsume "someone": <person#someone>, <life_form>, <causal_agent>, and <entity>. Thus, each of these four concepts receives $\frac{1}{4}$ of the frequency of "someone" in the data ($\frac{100}{4} = 25$ in the example).[5]

The overall frequency of a noun concept $ncpt$ is calculated as the sum of all word frequency portions which are mapped to it, i.e.

[5]Note that this is a simple example since "someone" is not ambiguous in WordNet. If the word in question is ambiguous (like "bank" mentioned above), then its frequency has to be equally divided among all concepts which subsume any sense of that word.

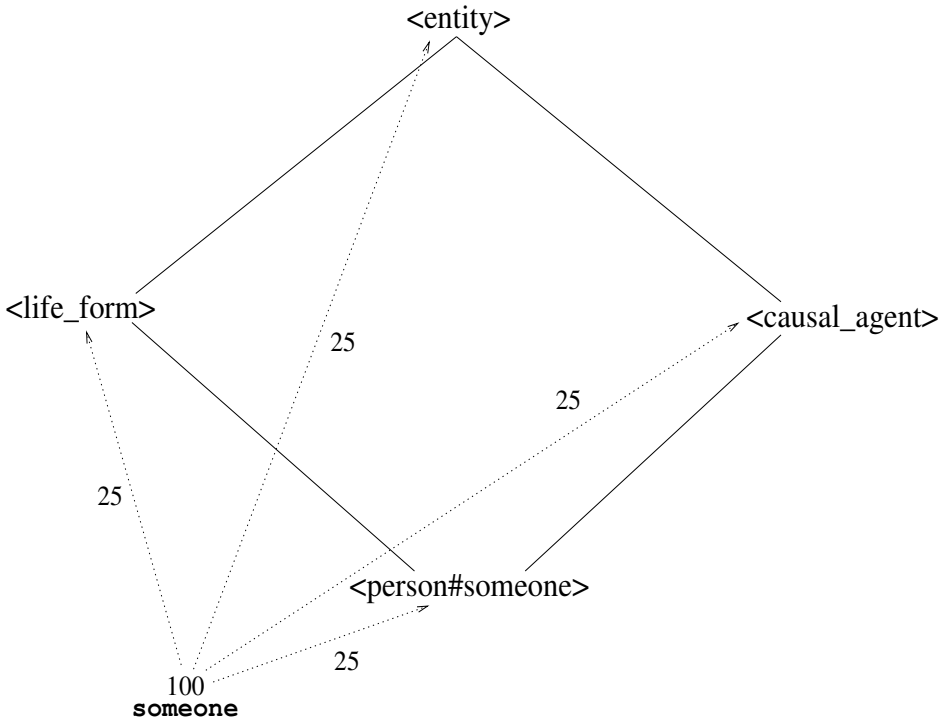


Figure 3.5: Frequency propagation by the word-to-concept approach

$$freq(ncpt) = \sum_{n \in words^{+}(ncpt)} \frac{1}{concepts(n)} \; freq(n) \tag{3.13}$$

where $freq(n)$ is the observed frequency in the data and $words^{+}(ncpt)$ is the set of words which are subsumed by $ncpt$, i.e. which are members of either the synset of $ncpt$ or the synset of one of its hyponyms. Henceforth, I will refer to this method as the *word-to-concept approach*, because it immediately maps a word form to all the concepts indicated by this word form.

The probability of a concept is obtained by maximum likelihood estimation:

$$p(ncpt) = \frac{freq(ncpt)}{N} \tag{3.14}$$

where $N = \sum_{ncpt'} freq(ncpt')$ is the size of the training data.

The joint frequency and the joint probability of a verb v and a noun concept $ncpt$ are computed analogously:

$$freq(v, ncpt) = \sum_{n \in words^{+}(ncpt)} \frac{1}{concepts(n)} \; freq(v, n) \tag{3.15}$$

$$p(v, ncpt) = \frac{freq(v, ncpt)}{N} \tag{3.16}$$

Similarly, the marginal probability of a verb form is

$$p(v) = \frac{freq(v)}{N} \tag{3.17}$$

These joint and marginal probabilities are required to calculate the conditional probabilities $p(ncpt|v)$, which, as we will see, are crucial for quantifying selectional preference.

The word-to-concept approach yields a probability distribution over all concepts in the hierarchy: the probabilities $p(ncpt)$ of all noun concepts sum to 1. The same holds for the conditional probabilities $p(ncpt|v)$.

Note that dividing the frequency count of a noun equally among each potentially corresponding concept in equation (3.13) introduces noise, since this results in assigning a certain amount of probability mass to concepts which correspond to incorrect senses of the noun instances in the training data. Resnik claims that the negative impact of this effect is rather low. He argues as follows: Due to its selectional preferences, a verb tends to co-occur with complement nouns which have senses complying with these preferences. The frequency portions assigned to these senses provide cumulative evidence (i.e. high frequency counts) for the correct concepts. In contrast, incorrectly assigned frequency portions tend to disperse throughout the hierarchy, only accumulating low frequency counts. (Ribas 1994) reports experiments which indicate that Resnik's assumption applies to a large extent, but that nonetheless the accumulation of incorrect senses has a significant impact on the acquired

preferences. The experimental results which I will describe in the chapters 5 and 7 exhibit a similar effect. I will return to this problem in section 5.2.

3.4.1.2 Selection and Information

In section 2.2.2, I pointed out that Jackendoff regards selectional restrictions as information which a verb supplies about its arguments. In a sense, the modelling of selectional preferences proposed in (Resnik 1993) forms a "statistical counterpart" of this view. For his approach, Resnik adopts ideas from that branch of science that conceptualises information quantitatively and relates this term to the concept of probability, namely *information theory*, a theory developed by Claude Shannon (cf. (Shannon 1948))[6].

Basically, information theory understands information as the content of a message which has to be transmitted from a sender to a receiver. As a consequence, it deals with coding information as efficiently as possible. In the framework of this discipline, information is usually coded in bits. The amount of information is measured by the number of bits needed to encode it. Of course, the applied coding scheme, i.e. the assignment of a unique bit sequence to each sign, is crucial in this respect. If one has to code a sequence of signs (in our case, nouns which occur as the complement of a certain verb in a corpus), the simplest way to do this would be to represent each sign by a bit sequence of uniform length. However, if the probabilities of the individual signs differ significantly, it is more efficient (with respect to data compression) to assign shorter bit sequences to more probable (and thus more frequent) signs and longer bit sequences to less probable (and less frequent) signs.

The correlation between the amount of information a message bears and the probability of that message can be illustrated intuitively by the following scene: Suppose that it is summer in Italy, and somebody tells someone else that it is hot outside. The amount of information of this message is quite low, since this is exactly the weather you would expect then. In other words, the probability of hot weather in Italy during that season is close to 1. In contrast, if the message said that it is snowing, then the amount of information would be so high that this message would be likely to appear in newscasts all over the world. The reason for this is that the probability of snow in Mediterranean summers is considerably low.

Concerning the "shape" of an efficient coding scheme, it can be shown that the shortest average code length is achieved by assigning $\lceil \log_2 \frac{1}{p(x)} \rceil$ bits to a sign x with probability $p(x)$ (cf. (Cover & Thomas 1991)). Thus, if a good estimation of the probability distribution underlying the occurrence of the signs is available, then one can develop an efficient coding scheme based on this estimation. For example, imagine a setting where messages have to be transmitted which consist of an arbitrary sequence of four signs: A, B, C, and D. These signs could e.g. represent the possible states of a machine, and a message could be a protocol of a sequence of these states which is necessary to supervise the processes performed by that machine. Imagine further that the probabilities of the different states are as follows:

[6] For a comprehensive introduction to information theory, cf. (Cover & Thomas 1991).

$$
\begin{aligned}
p(A) &= \frac{1}{2} \\
p(B) &= \frac{1}{4} \\
p(C) &= \frac{1}{8} \\
p(D) &= \frac{1}{8}
\end{aligned}
$$

Now the task is to specify a coding scheme that represents each sign by a unique codeword (i.e. a bit sequence) in a way that the average code length of a message (which consists of a sequence of codewords) is minimised. In order to avoid ambiguities, this coding scheme should be a so-called *prefix code* (or *prefix-free code*). This means that no codeword may be a prefix of any other codeword. A prefix code enables the receiver of a message to unambiguously determine where one codeword ends and the next one starts in the received bit stream. The optimal coding scheme should assign shorter codewords to signs with higher probability and longer codewords to signs with lower probability. Moreover, as stated above, the lengths of the codewords should correspond to the logarithm of the reciprocal of their probabilities. Thus, for the different signs in the example the following code lengths are appropriate:

$$
\begin{aligned}
A: &\quad \log_2 \tfrac{1}{p(A)} = \log_2(2) = 1 \\
B: &\quad \log_2 \tfrac{1}{p(B)} = \log_2(4) = 2 \\
C: &\quad \log_2 \tfrac{1}{p(C)} = \log_2(8) = 3 \\
D: &\quad \log_2 \tfrac{1}{p(D)} = \log_2(8) = 3
\end{aligned}
$$

Indeed, there is a coding scheme that satisfies these conditions:

$$
\begin{aligned}
A &\equiv 0 \\
B &\equiv 10 \\
C &\equiv 110 \\
D &\equiv 111
\end{aligned}
$$

Algorithms (e.g. *Huffman coding*) exist which, given a probability distribution over the signs to be encoded, yield a (nearly) optimal coding scheme.

Entropy is a measure of the expected amount of surprise that a message will bring about. Formally, entropy is the lower bound of the average code length needed to represent a value of a random variable X and is given by

$$
H(X) = -\sum_x p(x) \log\ p(x) = \sum_x p(x) \log \frac{1}{p(x)} \tag{3.18}
$$

with $\log$ base 2, due to binary coding.

Relative entropy (or *Kullback-Leibler distance*) is a measure of the distance between two probability distributions p and q. More exactly, it quantifies the cost (the number of additionally required bits) of assuming distribution q when the real distribution is p. Relative entropy is defined as

$$D(p\|q) = \sum_x p(x) \log \frac{p(x)}{q(x)} = \sum_x p(x)(\log \frac{1}{q(x)} - \log \frac{1}{p(x)}) \tag{3.19}$$

If one uses q to construct a code for a random variable X with probability distribution p, the average code length is $D(p\|q)$ bits higher than the entropy of X.

3.4.1.3 Preference Strength and Selectional Association

Resnik refers to the amount of information that a predicate provides for a certain argument as *selectional preference strength*. The selectional preference strength quantifies how strong the predicate constrains an argument semantically. For example, "eat" has a greater selectional preference strength for its object than "see". Formally, this quantity is expressed by the relative entropy between the probability distribution of noun concepts occurring in a certain grammatical relation to the verb in question and the probability distribution of noun concepts occurring in this relation *regardless of a particular verb*:

$$\eta_v = D(p(ncpt|v)\|p(ncpt)) = \sum_{ncpt} p(ncpt|v) \log \frac{p(ncpt|v)}{p(ncpt)} \tag{3.20}$$

Note that the grammatical relation r is kept implicit in equation (3.20). More precisely, this equation should be

$$\eta_{v,r} = D(p(ncpt|v,r)\|p(ncpt|r)) = \sum_{ncpt} p(ncpt|v,r) \log \frac{p(ncpt|v,r)}{p(ncpt|r)}$$

I adopted the notational simplification from Resnik and other authors mentioned in this chapter (e.g. Li and Abe) and will maintain it throughout this work for better readability. It is justified by the fact that in the approaches described here, each syntactic relation is taken into account separately. Anyway, the reader should keep in mind that the presented formulae which model the selectional association between verbs and nouns are related to a particular argument type.[7]

Equation (3.20) quantifies the cost of assuming that the distribution is $p(ncpt)$ whereas the real distribution is $p(ncpt|v)$, and thus "the cost of not taking the predicate into account. Therefore, in a very direct way, the selectional preference strength of a predicate can be understood as the amount of information that it carries about its argument." (Resnik 1993, p. 58) It should be pointed out that the preference strength defined in this way is guaranteed to be non-negative and equals to 0 if and only if $p(ncpt|v) = p(ncpt)$ for all $ncpt$. This is a property of relative entropy (cf. (Cover & Thomas 1991)).

[7] Ribas (1995*b*) also examines the possibility to replace $p(ncpt|r)$ by $p(ncpt)$, which in this case refers to a distribution derived from all nouns in the corpus. Of course, the simplified notation does not capture this distinction.

The overall preference strength is a measure which provides interesting characteristics about a verb. For example, Resnik found an empirical correlation between preference strength and argument omissibility: Verbs which exhibit a high preference strength for their object (e.g. "eat") allow object omission; verbs with a weak preference strength (e.g. "see") do not. In my view, this finding underlines the analogy between Resnik's model and Jackendoff's analysis, since it captures a correspondence between selectional constraints and incorporation, which Jackendoff treats in a uniform manner.

Nevertheless, preference strength is not the quantity we are primarily looking for, since it does not tell *which* concepts a verb prefers. Resnik quantifies the preference of a certain concept by a measure that he calls *selectional association*:

$$A(v, ncpt) = \frac{1}{\eta_v} p(ncpt|v) \log \frac{p(ncpt|v)}{p(ncpt)} \tag{3.21}$$

Note that this is the addend of the sum in equation (3.20) that corresponds to concept $ncpt$ (i.e. the contribution of $ncpt$ to the preferential profile of v), normalised by the overall preference strength η_v. This normalisation yields a preference value for an individual concept which is independent from the overall preference strength of the verb. This formula takes into account two things: the *absolute probability* of $ncpt$ occurring as complement of verb v and the *deviation* of this probability from the general probability of $ncpt$.

Resnik does not aim at finding a representative set of noun concepts which express the selectional preferences of a verb at an adequate level of abstraction. Rather, he stores the verb's preference values for each noun concept in the hierarchy. The advantage of this procedure is that the acquired information about preferences is completely available for NLP applications. For example, Resnik suggests a simple method which employs this information for word sense disambiguation: To disambiguate a noun which co-occurs with a certain verb, compare the respective preference values of all its senses *and their hyperonyms*, and select the sense that is related to the concept with the highest value. In this way, the acquired association scores allow the correct disambiguation of "baseball" in "hit some baseball" vs. "play some baseball". The former instance refers to a ball as physical object, while the latter instance refers to the game. Figure 3.6 lists concepts occurring at the object position of "hit" which subsume some sense of "baseball", together with their respective selectional association scores.

Among these concepts, <physical_object> has the highest preference value. Thus, the sense referring to a ball is selected here. As figure 3.6 shows, concepts with particularly high preference values are related to this sense, whereas concepts with particularly low preference values are related to the sense referring to the baseball game. For "play", the reverse is true. In particular, the concept with the highest association score (2.73) is <game>, which is related to the correct sense in this case.

3.4.2 Ribas

3.4.2.1 Mapping the Training Data to WordNet: the Word-to-Sense Approach

To map the noun forms in the training data to WordNet concepts, Ribas (cf. (Ribas 1995*a*), (Ribas 1995*b*)) uses a strategy which differs form Resnik's word-to-concept approach in the way of estimating concept frequencies. Resnik divides the frequency count of a noun n among all concepts $concepts(n)$ which subsume n, i.e. concepts whose synsets contain n *and* the hyperonyms of

$ncpt$	$A(hit, ncpt)$
<physical_object>	4.48
<artifact>	4.27
<entity>	2.25
<ball>	0.13
<game_equipment>	0.11
<equipment>	0.07
...	...
<sport>	-0.00
<game>	-0.00
<diversion>	-0.01
<contest>	-0.01
<competition>	-0.01
<group_action>	-0.27
<activity>	-0.34
<act>	-0.85

Figure 3.6: Concepts occurring as object of "hit" and subsuming some sense of "baseball", and the corresponding association scores

these concepts. Ribas, in contrast, divides this frequency count only among the concepts $senses(n)$ which represent a sense of n, i.e. those concepts whose synsets contain n. Henceforth, I will call this mapping strategy the *word-to-sense approach*. In section 3.4.1.1, I noted that a noun does not only provide evidence for its senses, but also for the hyperonyms of these senses. Ribas accounts for this fact by percolating the frequency counts obtained for noun senses upwards in the hierarchy. For example, "lamb" has two senses in WordNet, one referring to an animal and one referring to meat. Each of these senses receives a credit of $\frac{1}{2}$ for each occurrence of that noun in the data. The former sense has the hyperonyms <young_mammal>, <offspring>, <animal>, <life_form>, and <entity>. The hyperonyms of the latter sense are <meat>, <nourishment>, <food>, <substance>, <physical_object>, and <entity>. Each of these hyperonyms gets a credit of $\frac{1}{2}$ per occurrence and per sense it subsumes. Thus, <entity> gets a credit of 1 and the other hyperonyms a credit of $\frac{1}{2}$ for each occurrence of "lamb". The word-to-concept approach would give each sense and each hyperonym a portion of $\frac{1}{12}$ per "lamb" occurrence, since $|concepts(lamb)| = 12$.

Figure 3.7 takes up the example in figure 3.5, this time illustrating the word-to-sense approach. The frequency of "someone" (100) is mapped to the synset <person#someone>, which represents the corresponding word sense. (Recall that "someone" is monosemous in WordNet.) This count is completely propagated to all concepts that subsume <person>.

In general, the frequency of a concept is estimated as the sum of the frequency counts of the word senses which the concept subsumes. More formally, let $\nearrow$ denote the HAS_HYPERONYM relation and $\nearrow^*$ the reflexive and transitive closure of $\nearrow$. Furthermore, let $senses_{ncpt}(n) = \{nsns | nsns \in senses(n) \wedge nsns \nearrow^* ncpt\}$ be the set of senses of n which are subsumed by $ncpt$. Then, the frequency of a concept is estimated by the equation

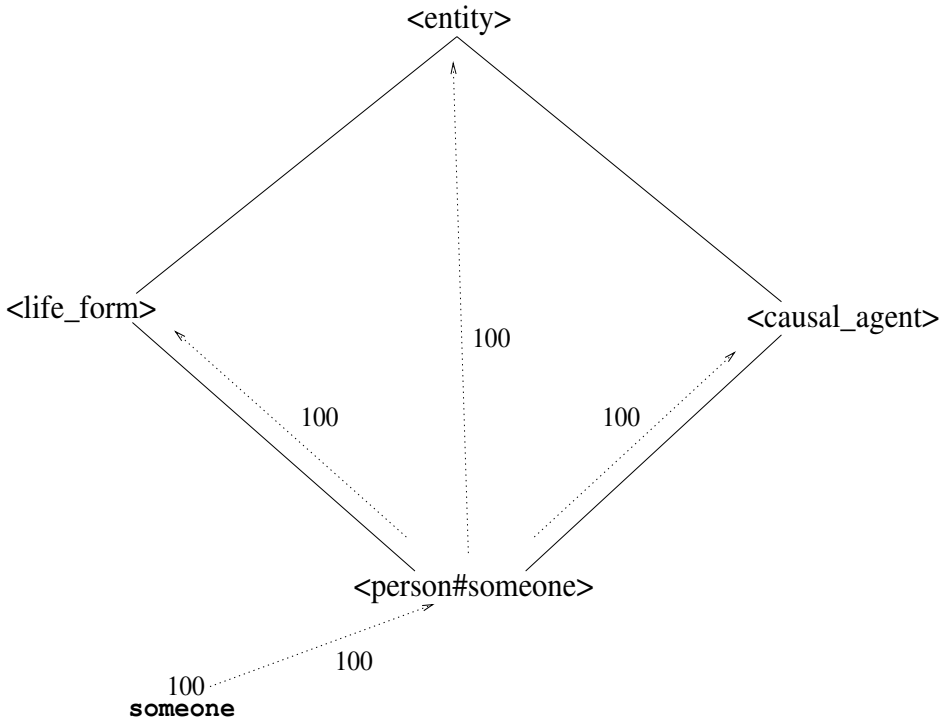


Figure 3.7: Frequency propagation by the word-to-sense approach

$$freq(ncpt) = \sum_{n \in words^{+}(ncpt)} \frac{|senses_{ncpt}(n)|}{|senses(n)|} \, freq(n) \tag{3.22}$$

These frequency counts are employed to acquire probabilities by MLE, i.e. in the same way as described in section 3.4.1.1.

The fundamental difference between the word-to-sense approach and the word-to-concept approach is that given a certain noun, the word-to-sense approach assumes a probability distribution over its senses:

$$\sum_{nsns \in senses(n)} p(nsns|n) = 1 \tag{3.23}$$

This approach views the WordNet hierarchy as an inventory of concepts with implication relations among each other. A hyponymy/hyperonymy relation between two concepts indicates that one concept (the hyponym) implies the other (the hyperonym). This means that a concept inherits all the probability mass of its hyponyms. In particular, since the root of the hierarchy is implied by all concepts, its probability is 1.[8]

The word-to-concept approach assumes a probability distribution over all the concepts a given noun belongs to:

$$\sum_{ncpt \in concepts(n)} p(ncpt|n) = 1 \tag{3.24}$$

This approach views the WordNet hierarchy as an unsorted pool of concepts which represent a smaller or larger set of nouns. In this model, hyponymy/hyperonymy relations between concepts indicate a common (sub)set of nouns providing evidence for these concepts. This model is required for quantities which are based on probability distributions over the whole inventory of concepts, like Resnik's overall preference strength (cf. equation (3.20) on page 67). A consequence of this model which might be somewhat counterintuitive is that the probability of the root concept is below 1. This is because probability mass is not completely inherited by, but equally divided among hyperonyms.

3.4.2.2 'KatzFodoresque' Selectional Preferences

Ribas (cf. (Ribas 1994), (Ribas 1995*a*), (Ribas 1995*b*)) basically adopts Resnik's selectional association measure—actually, he abandons the normalisation by the overall preference strength. But in contrast to Resnik, he does not keep all noun concepts, but extracts a "representative set" of concepts to model the preferential behaviour of a verb. (Resnik calls this kind of representation "more Katz-Fodoresque selectional restrictions in the form of weighted disjunctions" (Resnik 1997)). To induce this set, he uses a simple greedy algorithm, which can informally be stated as follows:

[8] WordNet does not have a singular root concept, but contains a small number of *unique beginners*, which are the most general concepts. According to the word-to-sense approach, the probability of these concepts sum to 1. It is straightforward to create a "virtual" root node if required.

grammatical relation	$ncpt$	$A(present, ncpt)$
subject	<causal_agent>	4.15
	<organization>	0.45
	<city#urban_center>	0.27
	<administrative_district>	0.26
	<life_form>	0.14
	<physical_object>	-0.01
object	<message#content>	0.57
	<psychological_feature>	0.46
	<creation>	0.31
	<state>	0.24
	<activity>	0.11
	<attribute>	0.08

Figure 3.8: Selectional preferences for the subject and the object of "present", acquired by Ribas' approach

1. Initially, put all noun concepts into the set of candidate concepts.[9]
2. Select that concept $ncpt$ which has the highest preference value (i.e. the highest selectional association measure) and insert it into the set of representative concepts.
3. Remove $ncpt$ and all its hyponyms and hyperonyms from the set of candidates.
4. Repeat step 2 and 3 until the set of candidates is empty.

In this way, Ribas yields a set of concepts which is non-redundant in the sense that the concepts are mutually disjoint; i.e. each noun sense occurring in the data is represented by one and only one concept in this set. To provide an example, figure 3.8 shows selectional preferences for the subject and the object of "present", acquired by Ribas' approach (cf. (Ribas 1995*a*)).

As emphasised in section 3.1, a crucial issue is the abstraction level of the extracted concepts. Ribas (1995*a*, p. 48) explicitly mentions the importance of acquiring concepts at an adequate generalisation level. To understand the rationale behind his approach of selecting concepts with respect to that goal, we have to look more closely at how Ribas calculates the preference values:

$$A(v, ncpt) = p(ncpt|v) \log \frac{p(ncpt|v)}{p(ncpt)} = p(ncpt|v) \log \frac{p(v, ncpt)}{p(v)p(ncpt)} \tag{3.25}$$

As one can see from equation (3.25), two factors determine the selectional association score. The first factor, $p(ncpt|v)$, grows when moving up the hierarchy. General concepts have a higher probability than specific concepts. The second factor turns out to be an information theoretic measure which is known as the *pointwise mutual information*:

[9] This is not exactly true. To eliminate noise, Ribas drops concepts whose frequency falls below a certain threshold.

$$I(v, ncpt) = \log \frac{p(v, ncpt)}{p(v)p(ncpt)} \tag{3.26}$$

This measure is widely used to detect collocations (e.g. in (Church & Hanks 1990)). It quantifies the association of v and $ncpt$ by the ratio of their joint probability according to the corpus and their expected joint probability under the assumption that v and $ncpt$ are independent, i.e. neither attract nor repel each other. Actually, this measure tends to favour more specific concepts, for the following reason: Due to their specificity, their usage contexts are rather restricted compared to general concepts. In particular, they tend to co-occur with a rather limited set of verbs so that their occurrence probability is highly dependent on the occurrence of these verbs. (Recall McCawley's "diagonalise"–"matrix" example.) Thus, the mutual information between these concepts and the verbs with which they typically co-occur will be high.

Altogether, the selectional association measure in equation (3.25) consists of one factor which favours general concepts, and one factor which advantages specific concepts. Ribas' concept selection algorithm is based on the assumption that these factors balance each other in a way that the concepts at the appropriate generalisation level receive the highest preference values and are thus selected by the above-mentioned algorithm.

Extracting a set of mutually disjoint concepts to represent selectional preferences has assets and drawbacks. One main advantage is storage and processing economy. WordNet 1.5 contains more than 60 000 noun concepts. To take into account all concepts, a corresponding number[10] of preference values have to be stored for each argument of each verb. It is obvious that storing only a comparably small set of concepts and preference values per verb and argument type reduces space requirements and processing efficiency dramatically. Furthermore, as I already said in section 3.1, for some NLP tasks it is crucial to encode selectional preferences at an appropriate generalisation level (e.g. semantic inferencing). And, of course, this applies as well for the representation of selectional restrictions in a "human-readable" form, which matters here, since EuroWordNet can be employed as a valuable resource for (cross-)linguistic research.

The main drawback is that dropping most of the concepts implies a loss of information which might be useful for certain NLP applications. For example, a set of concepts induced by the above-mentioned algorithm can be employed for word sense disambiguation in the same way as Resnik used the complete set of concepts (s.a.). The main difference is that each word sense is captured by one concept rather than several concepts. It is possible that just those concepts which provide the decisive evidence for the correct sense of a word are eliminated by the selection algorithm. This would be the case if those concepts had a descendant which has a higher preference value but which is not a hyperonym of the correct sense. It is an empirical question to what extent using all concepts enhances the performance of a certain NLP application. Unfortunately, the investigation of this issue goes beyond the scope of this thesis.

The mutual information component in the selectional association measure brings about an important property, namely a clear-cut distinction between preference and dispreference (cf. section 3.1). Equation (3.26) shows that if a verb v and a concept $ncpt$ attract each other, i.e. co-occur with a higher probability than it would be assuming independence, then $I(v, ncpt) > 0$. If, on the contrary, the two items repel each other, i.e. their joint frequency is below the expected value under the independence assumption, then $I(v, ncpt) < 0$. According to equation (3.25), this value is multiplied

[10] Of course, one could drop concepts which have no instance in the training corpus.

by the probability $p(ncpt|v)$ to obtain the selectional association score. This probability scales the absolute value of this score, but does not have an impact on its algebraic sign. Thus, we can draw an obvious borderline to distinguish preferred and dispreferred concepts: a positive association score indicates preference, a negative score dispreference. Note that this borderline also holds for Resnik's selectional association measure (equation (3.21)). As mentioned, Resnik additionally provides a normalisation by the overall preference strength. This normalisation does not influence the algebraic sign, since the preference strength is always non-negative.

3.4.3 Li and Abe

The approach described in this section provides a theoretically well-founded principle for finding an appropriate level of generalisation for representing selectional preferences. Therefore, this approach satisfies the second suitability criterion mentioned in section 3.1 better than all other approaches presented in this chapter. For this reason, I employed this method as a starting point for developing my approach for learning selectional preferences.

3.4.3.1 The Minimum Description Length Principle

As Ribas, Li and Abe (cf. (Abe & Li 1996), (Li & Abe 1998)) aim at representing selectional preferences as a set of concepts at an adequate generalisation level. However, their approach is not based on simple heuristics, but on a well-founded theoretical account, namely the *Minimum Description Length Principle (MDL)* developed by Jorma Rissanen (cf. (Rissanen 1989), (Rissanen & Ristad 1992)). This principle is motivated by information theory and rests on the assumption that learning corresponds to data compression: The better one knows which general principles underlie a given data sample, the better one can make use of them to encode this sample efficiently. For efficient data compression, it is crucial to identify a probability model which determines the distribution that underlies the data to be compressed, since this distribution constitutes an efficient coding scheme (cf. section 3.4.1.2). According to this rationale, to capture the information provided by a data sample, one has to encode

(1) the probability model that determines a coding scheme and

(2) the data themselves (by employing that coding scheme)

The code length needed to encode (1) is called *model description length*, the code length needed to encode (2) the *data description length*. The MDL principle states that the best probability model is that which achieves the highest data compression, i.e. which minimises the sum of model description length and data description length. In our case, the sample of data to be encoded consists of the noun tokens (more exactly, their senses) which appear in a certain syntactic relation of a certain verb v in the examined corpus. In the subsequent formulae, this sample will be referred to by S_v.

3.4.3.2 A Tree Cut Model

Li and Abe represent the selectional preferences of a verb by a *tree cut model*. Such a model provides a horizontal cut through the noun hierarchy tree, so that the concepts which are located along this cut

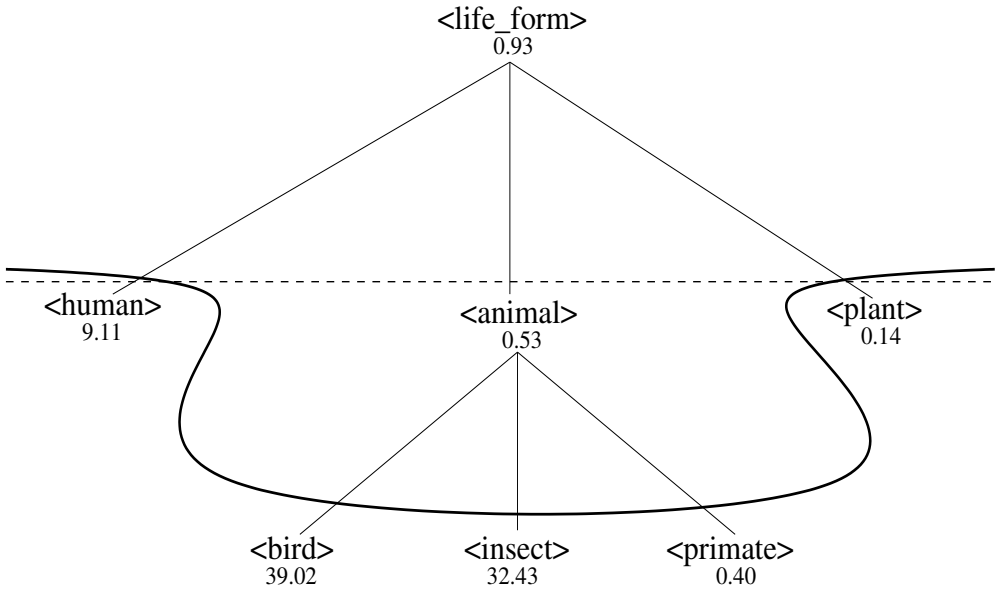


Figure 3.9: Two possible tree cut models for the subject of "fly"

form a partition of the noun senses covered by the hierarchy. A tree cut model consists of the concepts specified by a cut and the numerical preference values for these concepts. Figure 3.9 shows a portion of the WordNet hierarchy[11]—with preference values attached to the individual concepts—and two of the possible cuts across the hierarchy (indicated by a solid and a dashed line, respectively). Both of the corresponding tree cut models contain the concepts <human> and <plant>. However, while one of them contains the concept <animal>, the other one contains the more specific concepts <bird>, <insect>, and <primate>. This is an artificial example intended to illustrate plausible preference values (see below) and tree cut models for the subject of "fly".

In their different papers, Li and Abe propose two ways for calculating the preference values. In the following, I will concentrate on the alternative which is more suitable for the task of this thesis and shortly mention the other alternative at the end of this section. In (Abe & Li 1996), selectional preferences are quantified by a measure which Li an Abe call *association norm*:

$$A(v, ncpt) = \frac{p(ncpt, v)}{\hat{p}(ncpt)p(v)} = \frac{p(ncpt|v)}{\hat{p}(ncpt)} \tag{3.27}$$

[11] Strictly speaking, the hierarchy displayed here is a simplification. In the real WordNet taxonomy, <bird>, <insect>, and <primate> are not immediate hyponyms of <animal>, but occur at a lower level.

As a notational convention, I will designate probability distributions obtained by simple maximum likelihood estimation by $p(\ldots)$ and probability distributions estimated in a different way by $\hat{p}(\ldots)$. In some contexts, this distinction clarifies the description.

Note that this measure is closely related to the mutual information component in the preference value formulae of Resnik and Ribas (equations (3.21) on page 68 and (3.25) on page 72), the only difference being that the logarithm function is absent here. In principle, the association norm measures the association between a verb v and a noun concept $ncpt$ in the same way as mutual information, i.e. by the ratio of the joint probability and the product of the marginal probabilities of v and $ncpt$, which is equivalent to the ratio of the conditional probability of $ncpt$ given v and the probability of $ncpt$ regardless of a particular verb. In particular, this measure provides an obvious way to distinguish preferred and dispreferred concepts: A value above 1 indicates preference (v and $ncpt$ co-occur more frequently than by chance), a value below 1 dispreference (v and $ncpt$ co-occur less frequently than by chance).

A tree cut model represents a certain level of generalisation. It abstracts from the concepts below the cut. In the approach of Li and Abe, this is formalised by the stipulation that the preference value of a concept on the cut is inherited by its subconcepts. For example, the model indicated by the dashed line in figure 3.9 inherits the preference value for <animal> (which indicates dispreference) to its hyponyms. Thus, it abstracts from the fact that some of these hyponyms are preferred by "fly" (<bird>, <insect>), while others are dispreferred (<primate>). In contrast, the model indicated by the solid line captures these differentiations, since it generalises at a level below <animal>. Hence. the generalisation level represented by the solid-line model is more appropriate than the generalisation level expressed by the dashed-line model.

To select a tree cut model at the appropriate generalisation level, Li and Abe employ the Minimum Description Length principle (MDL) sketched above. A precondition for the applicability of MDL here is that a tree cut model determines a probability distribution $\hat{p}(nsns|v)$ over the senses of the nouns in the sample S_v (i.e. the senses of the nouns co-occurring with v); this probability distribution in turn constitutes a coding scheme. Actually, a tree cut model does determine such a distribution, provided that there is also an estimation of the marginal probabilities $\hat{p}(nsns)$ of noun senses regardless of a particular verb. These marginal probabilities can also be estimated on the basis of a similar kind of tree cut model by employing the MDL principle (cf. section 3.4.3.3). As noted above, the association norm of a concept is inherited by its descendants. A concept $nsns$ representing a sense of a noun in the data is subsumed by exactly one concept $ncpt_{nsns}$ on the cut (see below). Thus, $\hat{p}(nsns|v)$ can be obtained by

$$\hat{p}(nsns|v) = A(v, nsns)\hat{p}(nsns) = A(v, ncpt_{nsns})\hat{p}(nsns) \tag{3.28}$$

A probability estimation according to this equation (i.e. a unique substitution of $A(v, nsns)$ by $A(v, ncpt_{nsns})$) is only possible if for each noun sense there is a unique concept on the cut that subsumes this sense and thus represents it. In other words, a tree cut has to define a partition over the noun senses covered by the concept hierarchy. To ensure this, the structure of the hierarchy must exhibit two properties, which are basically not met by WordNet: First, the hierarchy must be a pure tree, i.e. all concepts (except the root) must have exactly one parent. This is necessary to guarantee that no noun sense is represented by multiple concepts on the cut (which would be the case if the cut contained two or more parents of a concept). Secondly, the noun senses must be represented by leaf nodes in the hierarchy, while the inner nodes represent more abstract concepts. This is required to ensure that all noun senses are below the cut and thus subsumed by a concept on it. As noted, the

WordNet noun hierarchy does not meet either condition. Firstly, it does not have a pure tree structure, but a DAG structure, i.e. some concepts have more than one immediate hyperonym. Secondly, as discussed in section 1.2.1, WordNet synsets simultaneously represent both word senses and abstract concepts: on the one hand, a synset models a particular sense of those words that are a member of it. On the other hand, a synset models an abstraction over those synsets that are descendants (hyponyms) of it. The tree cut approach requires that these two representational functions of WordNet nodes are separated as described above. Fortunately, there are several ways to transform the WordNet hierarchy in a way that makes it suitable for the tree cut approach. I will address this issue in detail in section 5.3.

To apply the MDL principle, it is necessary that the concepts constituting a tree cut model capture the complete probability mass of all the noun senses co-occurring with the examined verb. Otherwise this model could not properly determine a probability distribution over these noun senses. For this reason, Li and Abe employ the word-to-sense approach, since this approach completely propagates the frequencies (and thus the probability mass) of the noun senses at the leaves to the upward concepts in the hierarchy so that a concept which is located on the cut inherits the complete probability mass from its descendant leaves.

For selecting a tree cut model, the MDL principle is applied as follows: The number of bits $L(M)$ required to encode a sample S_v (i.e. a sample consisting of all noun senses co-occurring with the examined verb v) using a tree cut model M is given by

$$L(M) = L_{mod}(M) + L_{dat}(M) \tag{3.29}$$

with

$$L_{mod}(M) = L_{cut}(M) + L_{par}(M) \tag{3.30}$$

$L(M)$ is the sum of $L_{mod}(M)$, the model description length (the number of bits needed to encode the tree cut model M), and $L_{dat}(M)$, the data description length (the number of bits needed to encode the sample S_v by employing the coding scheme determined by M). Following the MDL principle, we search for the model M that minimises $L(M)$. The encoding of the tree cut model consists of two parts. The first part identifies the cut through the hierarchy. The code length needed for this part is $L_{cut}(M)$. The second part encodes the parameters of the model, i.e. the association norms (preference values) assigned to the concepts on the cut. The code length needed for that is $L_{par}(M)$. Thus, the model description length $L_{mod}(M)$ is the sum of these two code lengths.

For simplicity, it is assumed that all possible cuts through the hierarchy are equally probable. In section 3.4.1.2, I explained that the optimal code length for an item depends on the probability of that item. Therefore, if all possible cuts have the same probability, then the code length L_{cut} is constant for all cuts. For the task of minimising the description length, we can neglect this constant term. In other words, this task can be reduced to minimising the sum of the parameter description length $L_{par}(M)$ and the data description length $L_{dat}(M)$.

Encoding the parameters of the tree cut model amounts to encoding the association norm of each concept on the cut. Association norms are real numbers. This raises the question with which precision these numbers should be represented, i.e. how many bits should be spent to encode the decimal places of each association norm. This issue implies a trade-off: On the one hand, a higher precision leads

to a higher parameter description length, since more bits are required to encode the parameters. On the other hand, a higher precision reduces the data description length, because it results in a more precise probability estimation, which in turn yields a more accurate coding scheme. Conversely, a lower precision decreases the parameter description length, but increases the data description length. Li and Abe calculate $L_{par}(M)$ as:

$$L_{par}(M) = K\left(\frac{\log|S_v|}{2}\right) \tag{3.31}$$

K is the number of parameters in M (i.e. the number of concepts on the cut) and $|S_v|$ is the sample size. This formula means that for every concept on the cut, the association norm is represented by $\frac{\log|S_v|}{2}$ bits. This precision yields the optimum w.r.t. the trade-off just mentioned: it minimises $L(M)$ given a model M (cf. (Rissanen & Ristad 1992) for a proof).

The data description length is given by

$$L_{dat}(M) = -\sum_{nsns \in S_v} \log \hat{p}_M(nsns|v) \tag{3.32}$$

where $\hat{p}_M$ is the probability distribution over noun senses derived from M. This equation follows from the information-theoretic principles sketched in section 3.4.1.2: According to the optimal coding scheme determined by $\hat{p}_M$, each member $nsns$ of S_v is encoded by $-\log \hat{p}_M(nsns|v) = \log \frac{1}{\hat{p}_M(nsns|v)}$ bits. The sum of the codes for all these members yields $L_{dat}(M)$.

Overall, the rationale behind the tree cut approach is as follows: If the tree cut is located near the root, then the model contains only few concepts and the model description length will be low. However, the data description length will be high because the code for the data is based on the probability distribution determined by the concepts in the model, not on the real probability distribution of the noun senses. The coarser the model is, the more the corresponding distribution p_M deviates from the real distribution. And the more the supposed distribution deviates from the real one, the less efficient the coding scheme, and thus, the longer the code. On the other hand, if the tree cut is located near the leaves, the reverse is true: the fine-grained model fits the data well, resulting in an efficient coding scheme and thus a low data description length; but the large amount of concepts increases the model description length. Minimising the sum of these two description lengths yields a balance between compactness (expressing generalisation) and accuracy (fitting the data) of the model.

To acquire the tree cut model, Li and Abe perform a top-down search through the WordNet noun hierarchy. The search algorithm will be described in detail in section 3.4.3.4. Figure 3.10 shows a portion of the tree cut model Li and Abe acquired for the object of "buy".

3.4.3.3 Alternative Kinds of Preference Values

In (Li & Abe 1995) and (Li & Abe 1998), Li and Abe do not use the association norm of equation (3.27) but the plain posterior probability $p(ncpt|v)$ to quantify the preference of a concept $ncpt$ by a verb v. In these papers, they acquire a tree cut model which represents the preferences of a verb in an analogous way as sketched above. The difference is that the preference values assigned to the concepts on the cut are probabilities rather than association norms. The model description length can

$ncpt$	$A(buy, ncpt)$
<property>	3.97
<liability>	0.81
<asset>	3.38
<object>	2.15
<life_form>	0.73
<act>	0.88

Figure 3.10: Part of the tree cut model for the object of "buy", acquired following (Li and Abe 1996). The concepts are ordered due to their adjacency in the hierarchy.

$ncpt$	$p(ncpt\|buy)$
<asset>	0.10
<object>	0.30
<person>	0.02
<plant>	0.01
<action>	0.01
<activity>	0.02
<allotment>	0.03
<commerce>	0.01

Figure 3.11: Part of the tree cut model for the object of "buy", acquired following (Li & Abe 1995). The concepts are ordered due to their adjacency in the hierarchy.

be calculated as described above, since association norms and probabilities are encoded in the same way (as binary representations of real numbers). Regarding the data description length, Li and Abe stipulate that the probability of a concept on the cut is distributed equally among the senses which are subsumed by that concept. According to this assumption, a tree cut model determines a probability distribution over noun senses in the following way:

$$\hat{p}(nsns|v) = \frac{1}{|ncpt_{nsns}|}\, p(ncpt_{nsns}|v) \tag{3.33}$$

where $ncpt_{nsns}$ is the concept on the cut which subsumes $nsns$ and $|ncpt_{nsns}|$ is the number of noun senses which are subsumed by $ncpt_{nsns}$.

Figure 3.11 shows the portion of the tree cut model for the object of "buy" that corresponds to the portion displayed in figure 3.10, this time acquired by the alternative approach just sketched. (Both models originate from (Abe & Li 1996).)

The "probability tree cut approach" outlined here is employed for the "association norm tree cut approach" to estimate the marginal probabilities $\hat{p}(ncpt)$ which are required to determine association norms (cf. equation (3.27)) and noun sense probabilities (cf. equation (3.28)). The data sample used to estimate these marginals comprises all nouns which occur in a certain grammatical relation, regardless of a particular verb. Again, I refer the reader to section 3.4.3.4 for algorithmic details.

A comparison of the models in figure 3.10 and figure 3.11, which are based on the same training sample, illustrates that the two approaches yield different tree cuts. The cut in figure 3.11 is partially more specific. In particular, <person> and <plant> in figure 3.11 correspond to <life_form> in figure 3.10; likewise, <action>, <activity>, <allotment>, and <commerce> correspond to <act>. This difference reflects the fact that the two kinds of preference values model different ideas of preference. The probability $p(ncpt|v)$ models selectional preference as an *absolute* quantity: a concept which is more likely to co-occur with a verb receives a higher preference value than a concept whose co-occurrence is less likely. In contrast, the ratio $\frac{p(ncpt|v)}{p(ncpt)}$ models the preference of a concept $ncpt$ by a verb v *relative* to the overall occurrence rate of $ncpt$. In other words, this measure quantifies to what extent a verb attracts a concept over average or below average, respectively. Thus, if $p(ncpt|v)$ is rather low, but exceeds $p(ncpt)$ significantly, then $ncpt$ receives a high preference value. If, on the other hand, the co-occurrence of $ncpt$ and v is quite likely, but less likely than the overall occurrence of $ncpt$, then a low preference value is assigned.

To demonstrate the consequences of these different views of preference, let us assume two concepts $ncpt_1$ and $ncpt_2$ and a verb v with the following probabilities: $p(ncpt_1|v) = 0.002$, $p(ncpt_2|v) = 0.1$, $p(ncpt_1) = 0.001$, and $p(ncpt_2) = 0.2$. If one models selectional preference as an absolute quantity, the preference value of $ncpt_1$ is 0.002, which is one fiftieth of the preference value of $ncpt_2$ (0.1). If, however, one models selectional preference as a relative concept, the preference value of $ncpt_1$ is 2, which is four times as high as the value of $ncpt_2$, namely 0.5. One might object against the relative model that it seems counterintuitive to assign a higher preference value to that concept which co-occurs significantly less frequently with a verb. On the other hand, it is reasonable that the preference values reflect the peculiar impact which the verb in question imposes on the "semantic distribution" of its arguments. Some concepts a priori have a higher occurrence probability than others, just because they tend to be more salient in human communication (e.g. the concept <human> itself). The relative view of preference eliminates this a priori bias, while the absolute view maintains it. Both alternatives are intuitively justifiable.

With respect to the task of this thesis, the relative preference model has the advantage that it naturally distinguishes preference and dispreference (s.a.). In this way, it accounts for the categorical aspect of selection (cf. section 3.1). In the absolute model, the preference value is a probability, i.e. a number between 0 and 1. It is not straightforward to fix intervals of preference and dispreference within this range. A stipulation of such intervals is necessarily arbitrary.

3.4.3.4 The Acquisition Algorithm

This section provides a detailed description of the algorithm for acquiring tree cut models by means of the MDL principle to represent selectional preferences. This algorithm was essentially adopted from (Abe & Li 1996). I will discuss it in extenso because, as mentioned above, I have basically employed this approach for my work. In particular, I have implemented the algorithm presented here. It forms the core of my method for acquiring selectional preferences.

The data which form the input for the algorithm have been obtained from training data which consist of (v, n) pairs where v is a verb form and n is a noun form, and v and n are connected via a fixed grammatical relation (e.g. object). In a preprocessing step, the noun forms are mapped to the corresponding WordNet concepts by the word-to-sense approach (cf. section 3.4.2.1), i.e. the frequency of a noun is equally divided among all WordNet concepts which represent a sense of it; then the frequency counts of noun senses collected in this way are propagated upwards in the hierarchy to those

concepts that subsume the respective senses.

In the following, S denotes the sample of the senses of all nouns which occur in the data (regardless of a particular verb), S_v the sample of all noun senses which co-occur with a certain verb v, and t a node in the WordNet noun hierarchy (the latter notation reflects the property of a node to dominate a sub-tree in the hierarchy)[12].

The algorithm requires the following quantitative data:

- the frequency count $\sharp(t, S)$ of each concept t in the sample S (as obtained from the training data by the word-to-sense approach)
- the total count (size) $|S|$ of the sample S
- the cardinality $|t|$ of each concept t, i.e. the number of noun senses (leaves in the hierarchy) which are subsumed by t
- the frequency count $\sharp(t, S_v)$ of each concept t in the sample S_v (as obtained by the word-to-sense approach)
- the total count (size) $|S_v|$ of the sample S_v

The quantities $\sharp(t, S)$, $|S|$, and $|t|$ are required to retrieve a tree cut model that determines the marginal probabilities $\hat{p}(t)$ as sketched in section 3.4.3.3. These probabilities and the quantities $\sharp(t, S_v)$ and $|S_v|$ are required to retrieve the association norm tree cut model that represents the selectional preferences of v (cf. section 3.4.3.2). I will explain the details of their acquisition below.

Figures 3.12 and 3.14–3.16 contain a pseudo-code of the learning algorithm. As regards content, this pseudo-code is virtually identical to the code provided in (Abe & Li 1996). I make some algorithmic details and distinctions explicit which were implicitly given in their pseudo-code. For example, I notationally distinguish between, on the one hand, probabilities p and association norms A—which are single real numbers—and, on the other hand, tree cut models p_{cut} and A_{cut}—which comprise a set of concepts with associated probabilities or association norms. Furthermore, the code presented here reflects some details of my own implementation (which are compatible to the algorithm of Abe and Li).

Figure 3.12 shows the main procedure of the algorithm. This algorithm receives as input a set $verbs$ that contains the verbs for which selectional preferences shall be learned. The sample S is globally accessible (1.). The same applies to the root nodes of the hierarchy wn_roots (2.). As noted in section 3.4.2.1, the WordNet hierarchy does not have a singular root node. It rather has a small number of root nodes, i.e. very abstract concepts which do not have a hyperonym themselves. In the WordNet terminology, these nodes are called *unique beginners*. Figure 3.13 lists the unique beginners of WordNet 1.5. Of course, it is possible to "virtually add" an absolute root node to the WordNet hierarchy. Indeed, Li and Abe as well as McCarthy (2001) implemented this alternative. I decided not to do so, because an absolute root node by definition represents no information at all.

[12] To a large extent, the notation used in this section is borrowed from (Abe & Li 1996). This facilitates the comparison of my presentation with the presentation of Li and Abe. Furthermore, some minor notational errors in the pseudo-code presented by Li and Abe can straightforwardly be identified and corrected in this way.

algorithm Assoc-MDL($verbs$)
1. **global const** S
2. **global const** wn_roots
3. **global** $\hat{p}_{closure}$
4. **for each** $t_i \in wn\_roots$
5. γ_i := Find-MDL(t_i)
6. p_{cut} := append(γ_i)
7. **for each** $t_i \in wn\_roots$
8. Calc-p-Closure(t_i, p_{cut})
/* calculates $\hat{p}_{closure}$ */
9. **for each** $v \in verbs$
10. **for each** $t_i \in wn\_roots$
11. γ_i := Find-Assoc-MDL(v, t_i)
12. A_{cut} := append(γ_i)
13. output(A_{cut})

Figure 3.12: The main procedure of the basic algorithm to learn tree cut models

<phenomenon>
<possession>
<group#grouping>
<act#human_action#human_activity>
<event>
<state>
<shape#form>
<location>
<abstraction>
<psychological_feature>
<entity>

Figure 3.13: The unique beginners in the WordNet 1.5 noun hierarchy

sub-procedure Find-MDL(t)

1. $p(t) := \frac{\sharp(t,S)}{|S|}$
2. **if** t is a leaf node
3. **then** return($([t], p(t))$)
4. **else**
5. **for each** child t_i of t
6. γ_i := Find-MDL(t_i)
7. γ := append(γ_i)
8. **if** $\sharp(t,S)(-\log \frac{p(t)}{|t|}) + \frac{1}{2}\log|S| <$ $\sum_{t_i \in \gamma} \sharp(t_i,S)(-\log \frac{p(t_i)}{|t_i|}) + \frac{|\gamma|}{2}\log|S|$
9. **then** return($([t], p(t))$)
10. **else** return(γ)

Figure 3.14: The sub-procedure Find-MDL

The first part (4.–8.) of the algorithm (cf. figure 3.12) estimates the marginal probabilities $\hat{p}(t)$ based on the total sample S. This part has two sub-parts: Steps 4.–6. acquire a tree cut model p_{cut} that comprises a set of concepts t and associated probabilities $p(t)$. This is done by applying the sub-procedure Find-MDL separately to each root node in the hierarchy (i.e. <phenomenon>, <possession>, etc). This procedure processes the subtree dominated by the respective root and returns the best tree cut model (w.r.t. the MDL principle) for encoding the noun senses captured by that tree. Concatenating the cuts corresponding to the different unique beginners yields the model p_{cut} for the total hierarchy. This model determines the probability $\hat{p}(t)$ for *any* concept t. Therefore, steps 7.–8. employ p_{cut} to compute the probabilities for all concepts by the sub-procedure Calc-p-Closure, which again starts from each root node. The probabilities are stored in $\hat{p}_{closure}$. Since they are needed in the second part of the algorithm, $\hat{p}_{closure}$ is globally accessible (3.).

The second part of the algorithm (9.–13.) acquires for each verb v in $verbs$ (9.) the association norm tree cut model A_{cut} which represents the selectional preferences of v. Analogously to Find-MDL, the sub-procedure Find-Assoc-MDL acquires a cut for each subtree spanned by a root node (10.–11.); concatenating these cuts yields the complete model (12.), which is finally output (13.).

Figure 3.14 displays the sub-procedure Find-MDL. This procedure receives as input a concept t and acquires a "local" tree cut model for the subtree spanned by t. For example, when applied to <person>, Find-MDL obtains a tree cut model that contains concepts subsumed by <person>, i.e. concepts representing some kind of person. This model is acquired following the method proposed in (Li & Abe 1998) (cf. section 3.4.3.3). It comprises a cut through the respective subtree and the marginal probabilities of the concepts on that cut. The subtree dominated by t is processed top-down. First, the probability $p(t)$ is computed by simple MLE, i.e. by the ratio $\frac{\sharp(t,S)}{|S|}$ (1.). If t is a leaf node,[13] then we have the trivial case that the tree to be processed solely consists of the node t. In this case, the only possible tree cut model $([t], p(t))$ is returned (2.–3.). Otherwise (if t is an inner node), Find-MDL is recursively applied to each child of t and the resulting tree cut models are concatenated (5.–7.). Thus, the resulting model γ is the optimal solution among all possible cuts below t. For

[13]There are two different situations in which t is considered a leaf: first, t does not have hyponyms in the hierarchy; second, t has hyponym concepts, but these concepts do not cover any sense in S.

sub-procedure calc-p-closure(t, p_{cut})
1. **for each** child t_i of t
2. calc-p-closure(t_i, p_{cut})
3. **if** t is contained in p_{cut}
   /* $\hat{p}_{closure}(t)$ is directly defined by p_{cut} */
4. **then** $\hat{p}_{closure}(t) := p(t)$
5. **else if** t is above the cut defined by p_{cut}
6. **then** $\hat{p}_{closure}(t) := \sum_{t_i \in children(t)} \hat{p}_{closure}(t_i)$
7. **else** $\hat{p}_{closure}(t) := \frac{|t|}{|mother(t)|}\hat{p}_{closure}(mother(t))$

Figure 3.15: The sub-procedure Calc-p-Closure

example, for the subtree dominated by <life_form>, γ could comprise the immediate hyponyms (<person>, <plant>, <animal>,...), or concepts deeper in the hierarchy (<child>, <woman>, <flower>, <bird>, <insect>,...) or concepts from different abstraction levels (<person>, <plant>, <bird>, <insect>,...) There is only one possibility that has not been checked by the recursive calls of the procedure: the cut which just comprises the root t (in the example, <person>). Step 8. compares the code lengths needed to encode the noun senses covered by the subtree when employing the tree cut model $([t], p(t))$ or γ, respectively. Following equation (3.31) on page 78, the model description length is $\frac{1}{2}\log|S|$ for $([t], p(t))$ and $\frac{|\gamma|}{2}\log|S|$ for γ. As mentioned in section 3.4.3.3, the tree cut model determines a probability distribution over noun senses by equally dividing the probability mass of a concept on the cut among all senses which this concept dominates (cf. equation (3.33) on page 79). Thus, the probability of a noun sense dominated by a concept t_i on a cut is $\frac{p(t_i)}{|t_i|}$. (For example, suppose <bird> is part of the model with a probability of 0.07, and this concept subsumes 15 nodes which represent noun senses occurring in S. Then the probability for each of these senses is estimated as $\frac{0.07}{15}$.) Taking the negative logarithm to determine the corresponding code length for each instance (cf. section 3.4.1.2) and multiplying that code length by the number of instances in S covered by t_i yields a data description length of $\sharp(t_i, S)(-\log\frac{p(t_i)}{|t_i|})$ for each concept t_i on the cut. (For example, if the frequency of <bird> is 74, the code length for describing the senses subsumed by <bird> is $74 \times (-\log\frac{0.07}{15}$.) Summing the data description lengths corresponding to all concepts on each of the two cuts (this is trivial for $([t], p(t))$, which contains only one concept) yields the respective total data description lengths. Finally, that tree cut model which yields the smallest total code length (model description length + data description length) is returned by Find-MDL (9.–10.).

Figure 3.15 shows the sub-procedure Calc-p-Closure.[14] This procedure receives as input a concept t and the tree cut model p_{cut} acquired by Find-MDL. It employs this model to calculate the marginal probabilities $\hat{p}_{closure}$ of t and its hyponyms. In steps 1.–2., Calc-p-Closure is recursively applied to each child of t. Thus, $\hat{p}_{closure}$ is computed for all descendants of t. Steps 3.–7. calculate $\hat{p}_{closure}(t)$. There are three possible cases: If t is located on the cut defined by p_{cut}, then the corresponding probability $p_{cut}(t)$ is assigned to $\hat{p}_{closure}(t)$ (3.–4.). (For example, assume that <bird> is part of p_{cut} with the probability 0.07. Then this probability is adopted for <bird> in $\hat{p}_{closure}$.) If t is located above the cut defined by p_{cut} in the hierarchy, then the tree cut model does not affect the probability estimate for t. $\hat{p}_{closure}(t)$ can just be obtained by summing the probabilities of the children of t (5.–6.). This is

[14] This procedure is not explicitly part of the pseudo-code presented in (Abe & Li 1996). However, the inductive definition of $\hat{p}_{closure}$ implemented in 3.–7. is provided in that paper.

sub-procedure Find-Assoc-MDL(v, t)
/* v determines sample S_v */
1. $h(t|v) := \frac{\sharp(t,S_v)}{|S_v|}$
2. $A(v,t) := \frac{h(t|v)}{\hat{p}_{closure}(t)}$
3. **if** t is a leaf node
4. **then** return($([t], A(v,t))$)
5. **else let** $\tau := children(t)$
6. **for each** child $t_i \in \tau$ of t
7. $\gamma_i :=$ Find-Assoc-MDL(v, t_i)
8. $\gamma :=$ append(γ_i)
9. **if** $\sharp(t,S_v)(-\log A(v,t)) + \frac{1}{2}\log|S_v| <$
$\sum_{t_i \in \gamma} \sharp(t_i,S_v)(-\log A(v,t_i)) + \frac{|\gamma|}{2}\log|S_v|$
10. **then** return($([t], A(v,t))$)
11. **else** return(γ)

Figure 3.16: The sub-procedure Find-Assoc-MDL

equivalent to the upward propagation of frequency counts in the word-to-sense approach. Finally, if t is below the cut defined by p_{cut}, then $\hat{p}_{closure}(t)$ is determined by the tree cut model. As noted, the probability of a concept on the cut is equally divided among the senses (i.e. leaves of the hierarchy) which are captured by that concept. Each concept below the cut t_{bel} is a descendant of a certain concept on the cut t_{cut} and covers a subset of senses covered by that concept. Thus, the probability of t_{bel} is equal to the probability of t_{cut} times the proportion of senses that the descendant captures, i.e. $\frac{|t_{bel}|}{|t_{cut}|}$. (For example, suppose the concept <parrot> subsumes 2 nodes which represent senses occurring in the sample. The concept representing <parrot> in p_{cut}, <bird>, subsumes 15 sense nodes. Then the probability of <parrot> is estimated as $0.07 \times \frac{2}{15}$.) Note that this interrelation holds between t_{bel} and each of its ancestor concepts which are below or on the cut, since the probabilities of these concepts are defined accordingly. In particular, there is the following relationship between t_{bel} and its parent (immediate hyperonym):

$$\hat{p}_{closure}(t_{bel}) = \hat{p}_{closure}(parent(t_{bel}))\frac{|t_{bel}|}{|parent(t_{bel})|} \tag{3.34}$$

Step 7. calculates $\hat{p}_{closure}(t)$ according to equation (3.34).

Figure 3.16 displays the sub-procedure Find-Assoc-MDL. It gets as input a verb v and a concept t. It acquires a "local" association norm tree cut model (w.r.t. the sample S_v) through the subtree dominated by t. This procedure has a structure which is analogous to the structure of Find-MDL. In step 1., the probability $h(t|v)$ is calculated by MLE.[15] Step 2. computes $A(v,t)$ according to equation (3.27) on page 75, i.e. by dividing $h(t|v)$ by $\hat{p}_{closure}(t)$ (which, as noted above, is globally accessible). Note that it is this step which makes it necessary to acquire p_{cut} and $\hat{p}_{closure}$: Since every node in the hierarchy is processed by Find-Assoc-MDL, there must be a marginal probability estimate for every

[15]The notation $h(...)$, adopted from (Abe & Li 1996), obviously stands for *hypothesis*. $h(t|v)$ determines $A(v,t)$. Find-Assoc-MDL evaluates the hypothesis that $([t], A(v,t))$ is part of the optimal tree cut model for v.

concept. If t is a leaf, then the tree cut model $([t], A(v,t))$ is returned (3.–4.), else Find-Assoc-MDL is recursively applied to the children of t, yielding the tree cut model γ (6.–8.). Then, as in Find-MDL, the procedure compares the code lengths that $([t], A(v,t))$ and γ require for encoding the senses in S_v which are captured by t and its descendants (9.). (Here, on both sides of the inequation, the first addend calculates the data description length, the second addend the model description length.) A remark concerning the data description length is necessary here: This length is given by

$$\begin{aligned} L_{dat} &= \sum_{nsns} -\log h(nsns|v) && (3.35)\\ &= \sum_{nsns} -\log(A(v,t_{nsns}) \times \hat{p}_{closure}(nsns)) && (3.36)\\ &= \sum_{nsns} -\log(A(v,t_{nsns})) + \sum_{nsns} -\log \hat{p}_{closure}(nsns) && (3.37) \end{aligned}$$

Equation (3.35) corresponds to equation (3.32) on page 78; equation (3.36) follows from equation (3.28) on page 76 (t_{nsns} is the concept on the cut that represents $nsns$), and equation (3.37) follows from properties of the logarithm. As the second addend of this equation $\sum_{nsns} -\log \hat{p}_{closure}(nsns)$ does not change with the choice of A (and hence is constant for each tree cut model), this addend can be dropped when calculating and comparing the code lengths for the two cuts. Thus, in step 9. it is correct to add up the negative logarithms of the association norms $-\log A(v,t_i)$ for each concept t_i instead of employing the corresponding sense probabilities.

Finally, the tree cut model yielding the smallest code length is returned (10.–11.)

For the sake of completeness, I briefly mention a couple of minor notational errors in the pseudo-code in (Abe & Li 1996) and the corresponding pieces in the code presented here (the line numbers without parentheses refer to my pseudo-code, while those in parentheses refer to the code of Li and Abe):

- Find-MDL, line 8 (6): $\sum_{t_i \in \gamma}$ instead of $\sum_{t_i \in children(t)}$
- Find-Assoc-MDL, line 9 (9): $\sum_{t_i \in \gamma}$ instead of $\sum_{t_i \in \tau}$
- Find-Assoc-MDL, line 9 (9): $\frac{|\gamma|}{2}$ instead of $\frac{|\tau|}{2}$

Note that in Find-MDL and Find-Assoc-MDL, the selection of the appropriate cut is performed locally, i.e. the choice of a cut for a particular subtree does not depend on the choice of a cut for a different subtree. This local selection is possible because of the employed coding scheme: For the model description, (the parameter of) each concept on the cut is encoded separately.The identification of the cut itself is defined to have a constant code length for all possibilities, which does not affect the choice of the cut. Concerning the data description, since the hierarchy is required to be a pure tree with the word senses located at the leaves, it is guaranteed that two disjoint subtrees cover two disjoint sets of word senses. Thus, the decision which cut is optimal to encode the portion of the sample that is covered by a particular subtree can be made locally. Due to this locality property, the learning algorithm has polynomial complexity. More precisely, it is linear in the number of leaves of the hierarchy tree. Without the locality feature, the algorithm would have to consider every possible global cut, which would yield an exponential complexity.

3.4.4 Agirre and Martinez

Agirre & Martinez (2002) apply the word-to-concept approach to obtain frequencies and MLE probabilities. They extract their training data from SemCor (Miller et al. 1993), a corpus which is semantically tagged with WordNet senses. Thus, since they do not have to deal with the ambiguity of word forms, they only assign credit to the correct (i.e. the annotated) sense of a word instance and the hyperonyms of this sense.

As in (Li & Abe 1998), Agirre and Martinez quantify selectional preference by the conditional probability $\hat{p}(ncpt|v)$. The basic idea of their approach is that this probability is estimated by taking into account the conditional MLE probabilities of $ncpt$ *and all its hyperonyms* given v, instead of just adopting the MLE probability $p(ncpt|v)$. The motivation for this is to face the sparse data problem: for rare concepts, plain co-occurrence frequencies with a certain verb may not provide a reliable basis for estimating their preference values. In particular, it can be the case that a very specific concept which is preferred by a verb does not at all co-occur with that verb in the training data, just due to its generally infrequent usage. However, the hyperonyms of such a concept provide evidence for its preference (or dispreference). For example, it could be that the concept <chicken> never occurs as the object of "eat" in the data, but its hyperonym <food> occurs frequently at this position. In this case, <food> provides evidence for the preference of <chicken>.

Formally, Agirre and Martinez estimate $\hat{p}(ncpt|v)$ in the following way:[16]

$$\hat{p}(ncpt|v) = \sum_{ncpt \nearrow^* ncpt'} p(ncpt|ncpt') \times p(ncpt'|v) \tag{3.38}$$

This is the sum of the probabilities $p(ncpt'|v)$ for each $ncpt'$ subsuming $ncpt$. Each of these probabilities is weighted by the conditional probability of $ncpt$ given its hyperonym $ncpt'$, which is simply estimated by the ratio $\frac{freq(ncpt)}{freq(ncpt')}$. Note that this estimation requires that the frequencies are obtained by the word-to-concept approach. With the word-to-sense approach, the probabilities $p(ncpt'|v)$ of each hyperonym $ncpt'$ would comprise probability mass inherited from its subconcepts, e.g. the probability mass of $ncpt$ (which constitutes $p(ncpt|v)$). However, as equation (3.38) sums over all ancestors of $ncpt$, the same (inherited) probability mass would be added several times (though with differing weights). Thus, the word-to-sense approach would be inappropriate here. In contrast, with the word-to-concept approach frequency counts (and hence probability mass) is not inherited, but equally divided among $ncpt$ and its ancestors. This has the consequence that each addend in equation (3.38) corresponds to the portion of probability mass that $ncpt'$ receives due to the occurrence of $ncpt$ in the data. Thus, it is appropriate to add up all these portions to retrieve $\hat{p}(ncpt|v)$.

In section 3.1, I mentioned that techniques for learning selectional preferences of verb forms can easily be extended to acquire preferences of verb concepts as well. Agirre and Martinez propose two ways to do this. The first alternative is straightforward: Instead of employing probabilities conditioned by verb forms $p(ncpt'|v)$, they employ probabilities conditioned by the corresponding verb concepts $p(ncpt'|vsns)$. Since they use a corpus which is tagged with WordNet concepts, the latter probabilities can be obtained in the same way from the data as the former. Thus, the formula used for this alternative is analogous to equation (3.38):

[16] The corresponding formula in (Agirre & Martinez 2002) is misleading because the difference between $\hat{p}(ncpt|v)$ and $p(ncpt'|v)$ is not made explicit.

$$\hat{p}(ncpt|vsns) = \sum_{ncpt \nearrow^* ncpt'} p(ncpt|ncpt') \times p(ncpt'|vsns) \tag{3.39}$$

It would also be possible to apply this alternative when using a corpus which is not semantically annotated. This requires a mapping from verb forms to their senses, i.e. probability distributions of the form $p(vsns|v)$. Section 5.2 will address the estimation of these distributions.

The second way which Agirre and Martinez propose for learning preferences of a verb concept takes into account the ancestors of this verb concept in the same way as for noun concepts:

$$\hat{p}(ncpt|vcpt) = \sum_{ncpt' \nearrow^* ncpt} \sum_{vcpt' \nearrow^* vcpt} p(ncpt|ncpt') \times p(vcpt|vcpt') \times p(ncpt'|vcpt') \tag{3.40}$$

According to this approach, the selectional preferences of the hyperonyms of a verb concept provide evidence for the selectional preferences of that concept. Again, the rationale behind this is to overcome data sparseness. For example, if the concept <devour> never takes the object <chicken> in the corpus, but there is a high probability $p(\text{<chicken>}|\text{<ingest>})$, then equation (3.40) infers that <chicken> is preferred by <devour> to a certain extent, since <ingest> is a hyperonym of <devour>.

As Agirre and Martinez aim at integrating selectional preferences in WordNet, they have to apply a method to select a representative set of relations between verb and noun concepts in order to avoid redundancy and include only those links with the highest preference values. They employ a pruning algorithm which in essence is identical to Ribas' greedy algorithm for concept selection.[17] However, they propose an extension of this algorithm to $(vcpt, ncpt)$ pairs: From the candidate space, which initially comprises all pairs, select the pair $(vcpt, ncpt)$ with the highest preference value; remove all pairs $(vcpt', ncpt')$ from the candidate space where either $vcpt'$ subsumes $vcpt$ and $ncpt'$ subsumes $ncpt$, or $vcpt'$ is subsumed by $vcpt$ and $ncpt'$ is subsumed by $ncpt$; repeat these steps until the candidate space is empty.

Figure 3.17 shows selectional preferences for the object of two different senses of "know", acquired with and without generalising over verb hyperonyms, i.e. according to equation (3.39) and (3.40), respectively. The results for <have_sex#know> illustrate the potential of the approach to overcome data sparseness: As this concept does not occur in the training corpus, only the generalisation over its hyperonyms enables the acquisition of selectional preferences.

The strategy of interpolating concept probabilities by employing the probabilities of the concept's ancestors can be very helpful for certain NLP tasks. (Agirre & Martinez 2002) report WSD experiments in which the application of this mixture model, which accounts for unseen data, significantly improves recall. Regarding the task of this thesis, this approach is problematic if applied to verb concepts. To understand why this is the case, let me cite some interesting general considerations and

[17] One can justify the suitability of this algorithm w.r.t. the acquired generalisation level of selectional preferences in an analogous manner as for Ribas' approach. Like Ribas' association score, the formulae for the preference value proposed by (Agirre & Martinez 2002) contains one factor which favours specific noun concepts, namely $p(ncpt|ncpt')$, and one factor which favours general concepts, namely $p(ncpt'|vsns)$ or $p(ncpt'|vcpt')$, respectively. Thus, it is reasonable that the combination of these factors yields the highest preference values at the appropriate generalisation level.

	without verb concept generalisation		with verb concept generalisation	
$vcpt$	$ncpt$	$\hat{p}(ncpt\|vcpt)$	$ncpt$	$\hat{p}(ncpt\|vcpt)$
<know#cognise>	<communication>	0.1128	<abstraction>	0.1030
	<quantity#amount>	0.0615	<cognition#knowledge>	0.0386
	<attribute>	0.0535	<physical_object>	0.0333
	<physical_object>	0.0389	<act#human_action>	0.0198
	<cognition#knowledge>	0.0307	<group>	0.0183
<have_sex#know>			<person>	0.6798
			<egg>	0.1793
			<physical_object>	0.1254
			<body_part>	0.0193

Figure 3.17: Selectional preferences for the object of different senses of "know", with and without generalising over the verb hyperonyms

examples pointed out by Fellbaum (1990). These considerations address the correlation between the specificity of verbs in the WordNet hierarchy and the specificity of their selectional restrictions:

> As one descends in the verb hierarchy, the variety of nouns that the verbs on a given level can take as potential arguments decreases. This seems to be a function of the increasing elaboration and meaning specificity of the verb. Thus, *walk* can take a subject referring either to a person or an animal; most troponyms [i.e. hyponyms, AW] of *walk*, however, are restricted to human subjects. And *goose-stepping* is usually, though not necessarily, done by soldiers; this verb rarely takes *children* or *old people* as arguments. On the other hand, {*move, travel*} can take not only person or animal subjects, but also vehicles, or objects moved by external forces. Similarly, figures or pictures can *communicate* and *talk*, they can even *deceive* or *lie*, but they cannot *fib* or *perjure themselves*, as only human speakers can. A piece of news may *hit*, *touch* or even *grab* you, but it cannot *punch*, *stroke* or *collar* you; only people can be agents of these verbs. (Fellbaum 1990, p. 49)

Thus, predicting the preferences of a verb concept (e.g. <walk>) by using the preferences of its ancestors (e.g. <move#travel>) is likely to indicate noun concepts as preferred which are outside the specific range of preferences of the concept in question (e.g. <vehicle>). In figure 3.17, some of the preferences for <have_sex#know>, like <physical_object>, are inappropriately acquired because of this generalisation strategy. They are introduced by the hyperonyms <join> and <connect>.[18] After all, these concepts receive lower preference values than the correct concept <person>.

Therefore, acquiring selectional preferences by generalising over *verb concepts* may be helpful for NLP tasks for which coverage matters more than appropriate generalisation. For the task of this thesis, this possibility turns out to be inappropriate. In contrast, estimating the probability of *noun concepts* by taking into account their hyperonyms is not a priori problematic w.r.t. this task. Note that Li and Abe determine the probability of a noun concept below the cut by employing the probability

[18] <egg> is acquired due to a parsing error.

of its hyperonym on the cut. The crucial point is that the concept which determines the probability estimation of its subconcepts is at the appropriate generalisation level. The approach described in the following section explicitly addresses this issue.

3.4.5 Clark and Weir

The basic idea of the approach proposed by Clark & Weir (2002) estimates the probability $\hat{p}(ncpt|v)$ by means of a corresponding probability of one of its hyperonyms $ncpt'$. The rationale behind this is that a probability attributed to a concept which is higher in the hierarchy (and hence more frequent) can be estimated more reliably by MLE than a concept which is low in the hierarchy. (Clark and Weir employ the word-to-sense approach to retrieve frequency counts and MLE probabilities.) Of course, the probability $p(ncpt'|v)$ itself is not suitable to approximate $\hat{p}(ncpt|v)$, since in general, it is much higher. However, this does not apply for probabilities of the form $p(v|ncpt)$, which do not condition on the verb, but on the noun concept. For example, it is reasonable to assume that the probability p(run|<dog>) can be satisfactorily estimated by p(run|<canine>), or even by p(run|<animal>). Thus, if we have a concept $ncpt'$ that is a hyperonym of $ncpt$ and if $p(v|ncpt')$ is a good approximation of $p(v|ncpt)$, then we can use $p(v|ncpt')$ to estimate $\hat{p}(ncpt|v)$ in the following way:

$$\hat{p}(ncpt|v) = p(v|ncpt)\,\frac{p(ncpt)}{p(v)} \tag{3.41}$$

$$\approx p(v|ncpt')\,\frac{p(ncpt)}{p(v)} \tag{3.42}$$

(3.41) follows from Bayes' law. In (3.42), $p(v|ncpt)$ is replaced by $p(v|ncpt')$.

This approach requires to find for each concept $ncpt$, given a verb v, an appropriate hyperonym (henceforth indicated by $top(ncpt, v)$) which on the one hand is as general as possible and on the other hand is suitable to estimate $p(v|ncpt)$ as explained above. (For example, p(run|<entity>) would be a bad approximation for p(run|<dog>).) Clark and Weir consider a hyperonym $top(ncpt, v)$ as suitable if all concepts $ncpt'$ it subsumes have a similar probability $p(v|ncpt')$, which in particular means that $top(ncpt, v)$ solely subsumes concepts which behave comparably to $ncpt$ in this respect. It is easy to show that under these circumstances, $p(v|top(ncpt, v))$ is similar as well (cf. (Clark & Weir 2002, p. 190)). Thus, $top(ncpt, v)$ allows an adequate estimation of $p(v|ncpt)$.

To find the highest concept in the hierarchy which meets this condition, Clark and Weir employ an iterative bottom-up approach. This approach starts with $ncpt$ and moves up one level in the hierarchy after each iteration. Within an iteration step, the approach examines whether the children of the parent of the concept under consideration (i.e. its co-hyponyms) yield similar probabilities for the verb in question. If this is the case, then the parent concept becomes the concept under consideration in the next iteration step. If, however, the co-hyponyms of the concept under consideration behave inhomogeneously w.r.t. the verb, then the appropriate generalisation level is reached and the current concept is returned as $top(ncpt, v)$. For example, consider we want to find top(<dog>, run). The first iteration step examines the co-hyponyms of <dog>. It finds that the conditional probabilities of "run" given these co-hyponyms (e.g. p(run|<wolf>), p(run|<fox>), p(run|<bitch>) (and, of course, p(run|<dog>)) are similar. Thus, the algorithm moves one level up in the hierarchy. In the next step, it examines the hyperonym of <dog>, i.e. <canine>, and its co-hyponyms (e.g. <feline>,

$(ncpt, v)$	α	$top(ncpt, v)$
(<dream>,remember)	0.0005	<state>
	0.05	<state>
	0.5	<cognitive_state>
	0.995	<preoccupation>
(<man>,see)	0.0005	<animal>
	0.05	<mammal>
	0.5	<mammal>
	0.995	<man>
(<belief>,abandon)	0.0005	<psychological_feature>
	0.05	<mental_object>
	0.5	<belief>
	0.995	<belief>
(<coffee>,drink)	0.0005	<beverage>
	0.05	<beverage>
	0.5	<beverage>
	0.995	<beverage>

Figure 3.18: Generalising top concept $top(ncpt, v)$ for different $(ncpt, v)$ pairs and different α values

<bear>, etc.) and finds that the corresponding probabilities are still similar. In this way, further iteration steps are performed up to the point where <animal> and its co-hyponyms (e.g. <plant>) are examined. As p(run|<animal>) differs significantly from p(run|<plant>), the algorithm stops here and returns <animal> as top(<dog>, run).

To determine whether the children of a concept behave homogeneously or not, Clark and Weir use a chi-squared test where the assumption of homogeneity is the null hypothesis, which is rejected if there is significant evidence for inhomogeneity.[19]

As for any significance test, an important variable of the χ^2 test is the significance level α (alpha error), which is the probability that the null hypothesis is rejected although it is true. The smaller the value of α is chosen, the more conservative the is test, i.e. the more distinct evidence is required to reject the null hypothesis. In statistical science, α is set *before* the test is performed, in order to fix the level of conservativeness which is considered adequate. (Usual values are $\alpha = 0.05$ or $\alpha = 0.01$.) Clark and Weir, however, regard α as a variable whose appropriate value may be determined empirically to fine-tune the learning approach w.r.t. a specific application. The higher the alpha error, the easier the null hypothesis is rejected, and thus, the lower the generalisation level of $top(ncpt, v)$. This behaviour is illustrated by figure 3.18, which shows the generalising top concept $top(ncpt, v)$ acquired for different $(ncpt, v)$ pairs and different values of α. Note that for verbs with very strong preferences, e.g. "drink", the influence of α is low or even nonexistent.

Different choices of $top(ncpt, v)$ yield different estimates of $\hat{p}(ncpt|v)$. This has an impact on NLP applications employing these probabilities. Clark and Weir evaluate their approach by a pseudo-

[19]Technically, the contingency tables used for the χ^2 test are based on the joint frequency counts $freq(v, ncpt')$ which are used to estimate the corresponding probabilities $p(v|ncpt')$.

disambiguation task.[20] They report that for different values of α, the average number of generalised hierarchy levels differs significantly (from 4.5 for $\alpha = 0.0005$ to 1.9 for $\alpha = 0.995$). The differing amount of generalisation coincides with different disambiguation accuracies, with higher accuracy for higher values of α. For comparison, Clark and Weir applied this task to evaluate the approaches of Ribas and Li & Abe (1998). These approaches performed worse than the method of Clark and Weir. Their average generalisation was comparably high (4.1 and 4.7 levels, respectively). These results suggest that for this particular task, the optimal performance is achieved with a rather low (though not the lowest) level of generalisation.

The approach of Clark and Weir has several similarities with the approach of Li and Abe. Both methods aim at finding an appropriate level of generalisation for modelling selectional preferences. Furthermore, the probability $p(v|ncpt)$, over which Clark and Weir generalise, turns out to be closely related to the association norm proposed in (Abe & Li 1996):

$$p(v|ncpt) = \frac{p(ncpt|v)}{p(ncpt)}p(v) \tag{3.43}$$

$p(v)$ is constant for all concepts and does not make a difference w.r.t. generalisation. The ratio $\frac{p(ncpt|v)}{p(ncpt)}$ is equal to the association norm proposed by Li and Abe (cf. equation (3.27) on page 75). The two quantities play an analogous role in the two approaches: while $p(v|top(ncpt, v))$ approximates the corresponding probabilities of the hyponyms of $top(ncpt, v)$, the association norm $A(v, ncpt)$ of a concept $ncpt$ on the cut is inherited by its hyponyms.

On the other hand, there are striking differences between the two approaches. Li and Abe learn a *complete* tree cut model by means of a top-down algorithm. In contrast, the generalisation procedure of Clark and Weir is a bottom-up search which is applied *separately* for each noun sense in the data. Thus, it is not guaranteed that the collection of the acquired $top(ncpt, v)$ concepts yields a (redundancy-free) cut through the hierarchy.[21] Clark and Weir emphasise that they do not aim at acquiring such a cut. Nor do they aim at acquiring concepts which represent selectional preferences at a level of generalisation which corresponds to human intuition. In their approach, the purpose of generalisation is to facilitate more reliable probability estimations.

3.4.6 Abney and Light

In a sense, the approach proposed by Abney and Light (cf. (Abney & Light 1998), (Abney & Light 1999)) is distinct from the other approaches introduced here. Their goal is to provide an integrated and sound stochastic model of selectional preference. The approaches that I have discussed so far are more or less hybrid, i.e. a composition of different (partly heuristic, partly theoretically well-founded) strategies. For instance, in all of these approaches, a heuristic strategy of mapping word forms to concepts on the one hand and the actual algorithm for learning selectional preferences on

[20] For this task, triples of the form (v, n, v') are constructed so that the pairs (v', n) do not occur in the data, whereas the pairs (v, n) do. The (v, n) pairs are removed from the training set. The task is to determine which of the two verbs is more likely to co-occur with the noun. This pseudo-disambiguation task was applied as well by other researchers for evaluation (cf. (Pereira et al. 1993), (Rooth et al. 1998), (McCarthy 2001)).

[21] It even is not guaranteed that the probabilities $\hat{p}(ncpt|v)$ estimated according to (3.41) and (3.42) sum to 1. Therefore, Clark and Weir perform a normalisation to obtain a sound probability distribution.

the other hand are two distinct modules. In contrast, Abney and Light aim at devising a stochastic generation model that naturally unifies these modules.

To achieve this, Abney and Light present an idea which seems appealingly elegant: They interpret the WordNet noun hierarchy as the structure of a Hidden Markov Model: the concepts correspond to the states and the hyponymy relations to the transitions. Each state can emit those words which belong to the corresponding synset. Such an HMM models the selectional preferences of a certain verb v for a certain grammatical relation r. More precisely, the HMM *generates* nouns which co-occur with v (as argument r). To train the HMM, the noun instances occurring in triples containing v and r are extracted from the data.

Formally, a HMM of this kind is defined by

- a set of states $\{q_1, \ldots, q_n\}$ where each q_i corresponds to a noun concept
- a set of possible emissions $W \cup \{\epsilon\}$ where $W = \{w_1, \ldots, w_m\}$ comprises the words (nouns) covered by the hierarchy and ϵ denotes non-emission
- a matrix $A = \{a_{ij}\}$ where a value a_{ij} denotes the transition probability from state q_i to state q_j. This probability is nonzero only if q_j is an immediate hyponym of q_i.
- a matrix $B = \{b_j(k)\}$ where a value $b_j(k)$ denotes the emission probability of word w_k by state q_j. These probabilities are defined such that the states corresponding to the leaves of the hierarchy emit words (i.e. the emission probability of ϵ is 0), whereas states corresponding to inner nodes are non-emitting (i.e. the emission probability of ϵ is 1).
- the initial state distribution $\pi = \{\pi_i\}$. This distribution is defined such that the state corresponding to the root of the hierarchy is the only initial state (i.e. has the initial state probability 1). The leaves of the hierarchy are the final states.

The restriction that all and only the leaf states emit words (and are final states) has the consequence that an observation sequence generated by the HMM consists of exactly one word, emitted by a final state. This guarantees that the word probabilities estimated by the HMM sum to 1. However, in order to meet this constraint, the WordNet hierarchy needs to be restructured. As mentioned already, the inner nodes in WordNet represent abstract concepts (which subsume other concepts) as well as word senses. For example, the concept <person> represents both a concept that subsumes any kind of person and a particular sense of the word "person". These two notions have to be separated. To achieve this, each inner node is duplicated, and one of the duplicate nodes represents the inner concept, maintaining all hyponymy relations, while the other node represents the word sense corresponding to the original node. This node becomes an additional hyponym of the inner concept and does not have any hyponyms itself. With this modification, the requirement that all and only the leaves represent word senses (and thus emit word forms) is fulfilled.

In an HMM defined in this way, word forms, word senses, and concepts are combined in a principled way. No stipulative heuristic for mapping forms to concepts is necessary, since this mapping directly arises from the transitions and emissions constituting the HMM. Furthermore, such a model allows the estimation of different sorts of probabilities and other quantities in a straightforward way. For example, different senses of a word correspond to different state sequences in the HMM (starting at the root and ending at the leaf which emits the word). The probabilities of these state sequences can be easily computed and employed for word sense disambiguation.

$ncpt$	$A(break, ncpt)$
<object>	0.033223
<law>	0.020298
<law_of_nature>	0.020287
<substance>	0.018689
<ice>	0.016658
<solid>	0.016407
<guidance>	0.015359
<rule>	0.014416
<entity>	0.014345
<crystal>	0.014334

Figure 3.19: Selectional preferences for the object of "break", retrieved from the respective HMMs

The model also allows the estimation of concept probabilities of the form $p(ncpt|v)$ or $p(ncpt)$, which are the components of the formulae to calculate the preference values in the approaches described so far. The former kind of probabilities requires an HMM trained on the complements of a particular verb, while the latter requires an HMM trained on all nouns. To estimate such concept probabilities, the HMM has to be extended by adding to each leaf a single transition to the root. The resulting HMM turns out to be ergodic, i.e. each state is reachable form any other state. A consequence of this is that the probability of being in a certain state at a time t converges to a single value for $t \rightarrow \infty$ (cf. (Abney & Light 1998) for a proof of this claim and the way to determine these probabilities). This value gives the estimate of the probability of the corresponding concept. Figure 3.19 shows selectional preferences for the object of "break" (taken from (Abney & Light 1999)) obtained by estimating concept probabilities and then calculating the preference values following Resnik's approach (cf. equation (3.21) on page 68).[22]

Finally, an HMM defines a probability distribution over observation sequences, i.e. over words in this case. This probability distribution could be employed as a component of a sophisticated language model.

Abney and Light found out that unfortunately, training the HMM with the standard forward-backward algorithm does not yield an adequate model of selectional preferences. In particular, the forward-backward algorithm does not bias the parameters towards favouring the correct senses of ambiguous words, which is crucial for the task. Rather, if one trains the HMM with initial parameter settings which unavoidably reflect a mixture of senses for a word, the resulting HMM usually models such a mixture as well. Abney and Light state that this is a property of the EM algorithm, which "strongly prefers mixtures containing small amounts of many solutions over mixtures that are dominated by any one solution" (Abney & Light 1999, p. 4). To overcome this problem, they introduce a smoothing technique to favour senses which are close to other concepts in the hierarchy for which there is much evidence in the data. For example, if the data contain many words denoting some kind of food, then the food sense of "meat" is favoured over, say, the sense referring to the essential part of an idea. However, this smoothing technique introduces new problems so that further balancing strategies have

[22]These results are obtained by the modified HMM learning approach outlined below.

to be employed.[23]

Abney and Light evaluated their extended HMM learning approach by a WSD task, for which they used the same training and test data as in (Resnik 1997). They found that Resnik's approach outperformed theirs. Abney and Light characterise the performance of their approach as "disappointing" and conclude:

> One possible lesson is that EM itself is inappropriate for this problem. Despite the fact that it has become the default method for uncovering hidden structure in NLP problems, it essentially averages together many possible solutions. Possibly, a less linear method that eventually commits to one or another hypothesis about hidden structure may be more appropriate in this case. (Abney & Light 1999, p. 4)

Nevertheless, the work of Abney and Light provides interesting insights in the behaviour of the EM algorithm and the problem of learning selectional preferences. A particular reason for mentioning this work here is that its basic idea is closely related to the strategy I will propose for computing frequency counts of noun concepts (cf. section 5.3).

3.4.7 Summary

I have discussed several approaches for learning and representing selectional preferences by means of statistical corpus analysis and WordNet. In particular, I have addressed some fundamental properties of these approaches and their suitability for the task of this thesis. In this section, I will briefly recapitulate these properties in the light of the suitability criteria for the task of this thesis mentioned in section 3.1.

With the exception of the method of Abney and Light, all approaches discussed here require a mapping from word forms to WordNet concepts as a preprocessing step. Essentially, two strategies for this mapping have been proposed: the word-to-concept approach divides the frequency of a word form directly among all concepts which subsume (some sense of) this word form. This approach is employed by Resnik and Agirre/Martinez. The word-to-sense approach divides the frequency of a word form among its possible senses and inherits the concept frequencies obtained by this step to the hyperonyms of the respective concepts. This approach is used by Ribas, Li/Abe, and Clark/Weir. Which mapping strategy is appropriate depends on the peculiarities of the learning approach that makes use of it. Concerning the task of this thesis, none of the two strategies is a priori preferable over the other.

All approaches presented in this section express selectional preferences as WordNet noun concepts. This satisfies one part of suitability criterion 1. This criterion states that selectional preferences should be represented as relations between verb concepts and noun concepts. However, most of the approaches described here effectively acquire relations between verb forms and noun concepts. Only Agirre and Martinez learn selectional preferences for verb concepts rather than verb forms. However, the other approaches can be adapted to do so as well. This requires a mapping from verb forms to verb concepts, which can be performed by applying the same techniques as are applied for nouns.

[23] Another problem is that the HMM model penalises senses which correspond to a long path in the hierarchy, since longer paths involve more probabilities to be multiplied so that the product decreases. Abney and Light address this problem by a further balancing technique.

It is crucial for the task of this thesis to select a set of concepts which represent the selectional preferences of a verb at an appropriate level of generalisation (cf. suitability criterion 2). Resnik and Abney/Light do not determine such a set at all. Ribas and Agirre/Martinez obtain such a set by employing a simple greedy algorithm. Li/Abe use a theoretically well-founded approach, the MDL principle, to determine a representative set of concepts. Clark/Weir employ a generalisation procedure separately for each word sense to obtain more reliable probability estimations. They do not aim at finding an explicit representation of selectional preferences. Their approach, which employs a χ^2 test, provides the interesting possibility to influence the level of generalisation by the choice of the significance level α. This allows to fine-tune the learning method w.r.t. a particular application. Although an empirical determination of the value of α is anything but theoretically sound, it is an effective way of obtaining the generalisation level which is optimal for a particular NLP application. In section 3.1, I argued that this possibility is very important w.r.t. reusability.

The approach of Agirre/Martinez differs from the other methods in that it does not only generalise over noun concepts, but also over verb concepts. To acquire the preferences of a verb concept, this approach takes into account the noun concepts which co-occur with the hyperonyms of that verb concept. This generalisation increases the coverage of the acquired preferences, but may yield too general or even completely inappropriate preferences. Thus, it is not suitable for the task of this thesis.

Regarding suitability criterion 3, an important issue is the distinction between absolute and relative preference, which underlies the way the preference values are calculated. Absolute preference quantifies the probability that a noun concept co-occurs with a verb. Relative preference quantifies the ratio between this probability and the overall probability of the respective noun concept. Li & Abe (1998), Agirre/Martinez, and Clark/Weir measure absolute preference; Abe & Li (1996) measure relative preference. Resnik and Ribas employ a combination of these two, i.e. relative preference which is scaled by absolute preference. As relative preference provides a principled way to distinguish preferred and dispreferred concepts (which is required by criterion 3), whereas absolute preference does not provide such a distinction, relative preference is better suited for the task of this thesis.

To conclude, this brief review shows that the approach which is most appropriate for the task of this thesis is the method proposed by Abe & Li (1996). This approach satisfies best the suitability criteria mentioned at the beginning of this chapter: It represents preferences as WordNet concepts (criterion 1), aims at finding a representation at an appropriate generalisation level covering all nouns co-occurring with a verb (criterion 2), and provides a principled way to distinguish between preferred and dispreferred concepts (criterion 3). Basically, these criteria are also satisfied by Ribas' approach. However, while this approach determines the appropriate generalisation level by a simple ad-hoc strategy, Li and Abe employ a theoretically well-founded method for that task. In this respect, the approach of Li and Abe is superior. Therefore, I used this method as starting point for developing an approach which is suitable for learning thematic role relations.[24] It turns out that various modifications and extensions are necessary to yield satisfying performance. In the next chapter, I will provide a motivation and description of the modifications and extensions I propose.

[24]Nonetheless, I will also report experiments which empirically evaluate the suitability of Ribas' method for the task of this thesis, cf. section 5.5.

Chapter 4

Acquiring Selectional Preferences for Thematic Role Relations: The Basic Strategy

The two following chapters provide a detailed description and discussion of the approach that I propose for acquiring selectional preferences in such a manner that the results can be integrated in wordnets as thematic role relations. While this chapter concentrates on the general acquisition strategy, the next chapter will address issues concerning preprocessing and evaluation.

In section 3.4.7, I have argued that the approach of Abe & Li (1996) is better suited for that task than other approaches proposed in the literature. Therefore, I decided to employ this approach. However, it turned out that the general strategy developed by Li and Abe requires an elementary modification to be applicable for my work. This chapter motivates and explains this modification.

Section 4.1 addresses a fundamental shortcoming of the method of Li and Abe regarding the task of this thesis. Preliminary experiments with that method, which I carried out in an earlier stage of this work, showed that it does not behave satisfactorily concerning the generalisation levels of the acquired tree cuts. In section 4.1.1, I present some results of these experiments which illustrate this inadequacy. These results gave reason to generally considering the suitability of the Minimum Description Length (MDL) principle for learning tree cut models. In section 4.1.2, I argue that the observed inadequate generalisations are an immediate consequence of an intrinsic behaviour of the MDL principle. Furthermore, I exemplarily sketch two MDL approaches for different tasks to show that for these approaches (like for many other MDL applications in the literature), the same behaviour is perfectly appropriate. I conclude that the MDL principle in its strict form is inadequate for the class of tree cut models.

In section 4.2, I propose a modification of the Li and Abe approach, namely the introduction of a weighting factor. I show that although this modification deviates from the MDL principle, it is compliant to the framework of Bayesian reasoning. I report the second part of the tests described in section 4.1.1, which indicates that the weighting approach is more appropriate w.r.t. generalisation. More systematic experiments, which will be described in the next chapter (cf. section 5.5), confirm this result.

The considerations and findings in this chapter are not only important for the development of a method

to learn thematic role relations. They might as well be of interest as a case study concerning conditions and modalities for adequately applying the MDL principle. In this sense, they might provide a deeper understanding of the Minimum Description Length Principle as such. This rather general point is not the main issue of this thesis. Nevertheless, I think that the subtleties regarding MDL I came across are worth discussing, since they might be of significance for applying this approach to new tasks (i.e. tasks which have not been addressed in the MDL literature so far).

4.1 The Inadequacy of MDL

4.1.1 Preliminary Experiments

In this section, I report preliminary experiments which I performed in a rather early stage of the work described in this thesis. The intention of these experiments was to test the approach of Li and Abe in order to get an intuitive idea of how it works in practice and whether the acquired results exhibit interesting patterns or characteristics. In particular, I wanted to examine the generalisation level of the cuts learned for different verbs. These experiments have been reported in (Wagner 2000, p. 39–40).

4.1.1.1 Modifications and Extensions

In essence, I used the algorithm described in section 3.4.3.4 for the experiments. However, some modifications and extensions were necessary or useful.

Transforming WordNet Most importantly, the problem that WordNet does not meet the requirements for the structure of the semantic hierarchy mentioned in section 3.4.3.2 (pure tree structure; only leaves represent word senses) needs to be solved. In other words, the WordNet structure has to be virtually modified[1] to fulfil these criteria. To ensure that every word sense is represented by a leaf, I followed (McCarthy 1997) and introduced for every inner node an additional node, making this node a hyponym of the inner node. This new node—which is a leaf since it does not have any hyponyms itself—represents the sense corresponding to the words in the node's synset, while the original inner node represents the semantic concept subsuming these senses and the senses of the node's hyponyms. (I will indicate the additional nodes by the prefix 'REST::' as they represent the "rest" of concept instances which is not captured by the subconcepts.) Recall that this solution was also employed by Abney and Light (cf. section 3.4.6). To transform the WordNet DAG structure into a tree structure, I virtually duplicated every node (and its descendants) which has more than one parent. In this way, the DAG is "broken into a tree". This solution was adopted from Li and Abe. How this duplication works will be described in section 5.3. Furthermore, I will discuss alternative solutions to the transformation issue proposed in the literature, sketch drawbacks of the different alternatives, and propose a more principled approach in that section.

Optimising Parameter Description There is a straightforward optimisation of encoding the tree cut model parameters. If a concept in a tree cut model does not have any instances in the data sample

[1] Here, *virtually* means that this modification is performed on runtime and not stored persistently.

S, it receives the association norm 0. It does not make sense to represent the parameter value 0 with $\frac{\log|S|}{2}$ bits. A more efficient coding strategy is to mark the concepts that have a non-zero parameter and represent the parameter values for those concepts only. First you need K bits, one for each concept on the cut, which indicate whether a concept occurs in the sample (1) or not (0). Then you need $\frac{\log|S|}{2}$ bits for every class that occurs in the sample. Thus,

$$L_{par}(M) = K + K_S\left(\frac{\log|S|}{2}\right) \tag{4.1}$$

K_S is the number of classes that have instances in the sample S. With this modification, one saves $(K - K_S)\frac{\log|S|}{2} - K$ bits. I calculated the model description length using equation (4.1) instead of equation (3.31) on page 78. In the pseudo-code presented in section 3.4.3.4, step 8 in Find-MDL and step 9 in Find-Assoc-MDL have to be changed accordingly.

Threshold To eliminate noise, I introduced a threshold in the following way: The algorithm compares possible cuts by traversing the hierarchy top down. If a class with a probability below 0.05 is encountered, then the traversal stops, i.e. the descendants of that class are not examined. This has also the advantage of limiting the search space. Li and Abe as well as McCarthy employ thresholding in some form.

4.1.1.2 Setting

To test the behaviour of the Li and Abe approach, I applied it to acquire selectional preferences for the direct object relation. I extracted verb–object instances from a portion of the British National Corpus (BNC)—parts A–E (about 40 million words)—with Steven Abney's CASS parser (cf. (Abney 1996*a*)).[2] This resulted in a sample of approx. 2 million verb–noun pairs. I chose the object relation for two reasons. First, in the literature this relation usually has been chosen to inspect and illustrate selectional preferences acquired by a certain approach. Recall that the examples shown in the last chapter (which have been adopted from the respective papers) all comprise preferences for the direct object. Second, the verb–object relation can be extracted by CASS with the necessary reliability.[3]

Then I applied the learning algorithm to calculate the selectional preferences of 24 test verbs and manually inspected the results. These include very frequent verbs with rather weak preferences, like "have", "be", or "take" (these verbs are compatible to almost any noun concept), rather infrequent verbs with strong preferences for particular noun concepts, like, "kiss", "drink", or "eat", as well as verbs which lie between these extreme cases, like "attack" or "defend". In particular, the verbs "kill", "murder", and "assassinate" were included because these verbs prefer nouns from the same domain but with different specificity: As (Ikegami 1993) points out, "kill" selects any kind of life form. In contrast, "murder" prefers human beings. The usage context of "assassinate" is even more restricted, since the object (Patient) of this verb usually refers to a person who is important in some sense. In the

[2] I thank Steven Abney and Marc Light, who made these data as well as their source code for acquiring selectional preferences available to me. Those parts of my implementation which deal with collecting and storing co-occurrence statistics from the sample make use of their code.

[3] Steven Abney (personal communication) reported 94.6% precision (61.4% recall) for objects as opposed to 78.3% precision (79.7% recall) for subjects. For our purposes, high precision is crucial and more important than high recall.

initial experiments described here, I followed the general practice to acquire selectional preferences for verb forms rather than verb concepts.

4.1.1.3 Results

The experiment revealed a significant drawback of employing the MDL principle for the task of this thesis. It turned out that the frequency of the examined verb in the sample has an undesirable impact on the generalisation level of the tree cut model: The algorithm tends to over-generalise (acquire a tree cut with few general concepts) for infrequent verbs and to under-generalise (acquire a tree cut with many specific concepts) for frequent verbs. As an illustrative example of that behaviour, let us look at the tree cuts for "kill" (frequency count 2820.68[4]), "murder" (frequency count 500.08), and "assassinate" (frequency count 54.48).

Figure 4.1 displays part of the tree cut model acquired for "kill". The cut is located below <life_form>. It includes <person>, <plant>, and the immediate hyponyms of <animal> like <domestic_animal>, <pet>, <chordate>, or <invertebrate>. It is remarkable that all these concepts are preferred (preference value > 1), except for <plant> (preference value 0.48). This is an interesting empirical finding that deviates from Ikegami's (1993) claim that any kind of life form is selected by "kill", which is intuitively plausible at first glance. This finding suggests that it is not common to refer to the action of causing a plant to stop living by the expression "to kill a plant". This example makes clear that a statistical model of selectional preferences primarily captures the *probable* rather than the *possible* complements of a predicate. Anyway, one could argue that a cut located at the concepts <person>, <animal>, and <plant> is more appropriate here than a cut at <life_form>. However, a cut below <animal>, like in figure 4.1 clearly is an under-generalisation, since it does not capture the fact that any kind of animals is preferred by "kill".

Figure 4.2 shows part of the tree cut model acquired for "murder". In contrast to the model for "kill", this cut is too general. It includes <life_form>, whereas "murder" specifically prefers objects referring to humans. For the verb "assassinate", the over-generalisation of the acquired cut is even worse. Figure 4.3 shows that this cut is located at <entity>, one of the root nodes in WordNet. However, the selectional preferences of that verb ("important persons") are even more specific than the preferences of "murder". The comparison of these verbs illustrates a general tendency of the Li and Abe approach: the lower the frequency of the examined verb, the more general the learned tree cut model. This tendency results in too general cuts for infrequent and too specific cuts for frequent verbs.

4.1.2 The Tree Cut Model Class and the MDL Principle

It would be an error to ascribe the behaviour of the tree cut approach w.r.t. generalisation to a coincidence of peripheral circumstances. Rather, this behaviour immediately follows from the MDL principle. As explained in section 3.4.3.2, this principle selects that model which minimises the sum of the code length L_{mod} needed to encode the model itself and the code length L_{dat} needed to encode the data by employing the coding scheme determined by the model. The more specific a tree

[4]Note that concept frequency counts are not necessarily cardinal numbers. According to the word-to-sense approach, these counts are the sum of the counts of the word senses which they subsume. The word sense counts, in turn, are generally no cardinals, because they are estimated by dividing frequencies of word forms by the number of senses of these forms.

$ncpt$	$A(kill, ncpt)$
...	...
<person#individual#someone#mortal#human#soul>	3.39
<domestic_animal>	30.89
<pet>	30.89
<female>	42.77
<male>	5.52
<adult>	61.79
<young#offspring>	11.65
<giant>	30.89
<survivor>	61.79
<herbivore>	30.89
<embryo>	6.18
<chordate>	7.32
<invertebrate>	8.97
<predator#predatory_animal>	15.45
<prey#quarry>	170.41
<REST::animal#animate_being#beast#brute>	289.00
<plant#flora#plant_life>	0.48
<microorganism>	31.21
<parasite>	62.41
<mutant#mutation#sport>	3.69
<nonvascular_organism>	18.42
<REST::life_form#organism#being#living_thing>	9.57
...	...

Figure 4.1: Part of the tree cut model for “kill”

$ncpt$	$A(murder, ncpt)$
...	...
<life_form#organism#being#living_thing>	4.01
...	...

Figure 4.2: Part of the tree cut model for “murder”

$ncpt$	$A(assassinate, ncpt)$
...	...
<entity>	2.05
...	...

Figure 4.3: Part of the tree cut model for “assassinate”

cut model, the better the probability distribution over the examined sample fits this sample, and thus the more efficiently the coding scheme determined by this distribution can encode the sample. On the other hand, a specific tree cut model contains more concepts than a general one, which means that more parameters (association norms) have to be encoded to describe it. Hence, the more specific the tree cut, the lower the data description length L_{dat}, but the higher the model description length L_{mod}.

An immediate consequence of this is a correlation between the frequency of the examined verb and the specificity of the acquired tree cut. If the verb under consideration has a high frequency, then a large amount of data (the verb's complements) has to be described. In this case, the total description length ($L_{mod} + L_{dat}$) is largely dominated by L_{dat}. Thus, the selection of a more specific tree cut model results in a considerable reduction of the data description length (due to the more efficient coding scheme), while the expense for encoding more parameters has a negligible impact on the total description length. In other words, the gain of a complex model for encoding the data outweighs the model cost. If, in contrast, the examined verb has a low frequency, then only few data have to be encoded. In this case, L_{dat} is rather small and the relative contribution of L_{mod} to the total description length is much more significant. Thus, the cost of encoding a specific model with many parameters outweighs the gain for efficiently encoding the data, since this gain is limited due to the small amount of data to be encoded. In summary, the higher the frequency of a verb, the higher the tendency of the Li and Abe approach to acquire a specific tree cut model.

However, this is not the desired behaviour. Generalisation should not be triggered by the sample size, but by the "semantic variety" of the instances in the sample: Nouns like "apple", "pear", or "strawberry" should generalise to <fruit>. Further instances like "pork" or "cake" should trigger generalisation to <food>, and yet further instances like "house" or "vessel" to <physical_object>. Interestingly, the comparison of the verbs "kill", "murder", and "assassinate" suggests the tendency that the frequency of a verb coincides with the semantic variety of its complements. Verbs with a higher frequency tend to occur in a higher variety of contexts than verbs with a lower frequency. For example, while "kill" co-occurs with humans or animals, "assassinate" only co-occurs with a certain sort of humans. The reason for this is the higher meaning specificity of "assassinate" (cf. Fellbaum's quotation in section 3.4.4). However, the meaning specificity of a word is one factor which influences its frequency, since words with a specific meaning only co-occur in restricted contexts. Of course, there are other factors which influence the frequency of a word (e.g. stylistic issues). However, *if* there is a correlation between the frequency of a verb and the generalisation level of its selectional preferences, then it should be opposite to the tendency of the Li and Abe approach described above: higher verb frequency should coincide with more general preferences, not with more specific ones.

This mismatch between the actual behaviour of the tree cut approach and its desired behaviour for learning selectional preferences cannot be resolved within the MDL paradigm. The bias towards favouring complex models grows with the size of the data sample is an inherent property of the MDL principle. Rissanen & Ristad (1992, p. 163f.) summarise the rationale behind the MDL principle as follows: "The MDL formula states that simpler models are to be preferred to more complex models unless a more complex model provides a significantly more compact description of the observables." If a more complex model yields a more compact coding of a certain data item (or, more generally, a certain pattern of data items) in the sample than a simpler model, then, of course, this results in a reduction of L_{dat}. The extent of this reduction is the larger the more frequently the data item (or the pattern) occurs in the data. Since this applies to each data unit in the sample, the selection of a more elaborate model brings about a significantly higher reduction of the data description length if the sample is large than if it is small. As de Marcken (1996) puts it:

> "Parameters that are only infrequently true must introduce substantial savings to be worth including [in a model, AW]; parameters with a widespread usage are beneficial even if they introduce only incremental improvements to the statistical model [...] The length of a [parameter] representation is independent of the number of times a parameter is used, so given enough evidence the benefits of any parameter will outweigh its costs." (de Marcken 1996, p. 51)

The more frequently a parameter is "used" (i.e. employed for encoding data items)[5], the more likely it is that the savings that its introduction causes for the encoding of the data outweigh the cost of its description. Thus, if the introduction of an additional parameter yields only little compression of the code for a *single* data item, but this parameter is used for encoding *many instances* of such data items, then its introduction is justified. And, of course, the "chance" of encoding enough data to outweigh the model cost grows with the size of the data sample to be encoded. Therefore, the MDL principle favours complex models for large samples and simple models for small samples. A large amount of data makes a complex model "affordable".

As noted above, this inherent characteristic of the MDL framework has undesirable effects for the task of learning tree cut models which represent selectional preferences. However, for other tasks for which the application of MDL has been proposed, this property is perfectly appropriate. The reason for this is that these tasks make use of different kinds of models. Their model classes differ from the class of tree cut models in the general relationship between the complexity of a model and its *conservativity*. (I use this term in the following way: A model is more conservative than another model if it makes weaker claims about the general regularities underlying the data sample under consideration.) To explain this statement, I will exemplarily discuss two tasks for which the MDL principle is employed and compare the underlying model classes to the class of tree cut models.

First, I discuss a task that is commonly used for illustrating machine learning approaches (e.g., cf. (Mitchell 1997)): the induction of decision trees. Quinlan & Rivest (1989) propose to apply the MDL principle for that task. A data sample from which decision trees can be learned comprises a collection of objects which are characterised by certain properties that are expressed as values of a fixed set of attributes. Furthermore, each object belongs to a specific class, which partly (in the ideal case, completely) depends on the object's properties. Figure 4.4 shows such a sample (adopted from (Quinlan & Rivest 1989)). Here, the objects are Saturday mornings which are characterised by the attributes Outlook (sunny, overcast, or rain), Temperature (hot, mild, or cool), Humidity (high or normal), and Windy (false or true). These Saturday mornings are classified according to whether they are suitable for some "unspecified activity" (P = positive) or not (N = negative).

A *decision tree* models the dependencies between the properties of the objects and their respective classifications. Each inner node of a decision tree is associated with a test on the values of attributes. For each possible outcome of this test, there is a branch from the node to a different node. In (Quinlan & Rivest 1989), such a test evaluates a single attribute. Each possible value is represented by a branch from the respective node. The leaves of a decision tree correspond to subsets of the objects in the sample. A leaf represents those objects which pass all tests on the path from the root node to that leaf. Thus, a leaf can represent more than one object, but an object cannot be represented by more than one leaf node. Additionally, each leaf is associated with a particular class label which classifies those data items which that leaf represents. A *perfect decision tree* has the property that all objects which

[5]For example, a parameter of a tree cut model—an association norm of a concept on the cut—is used to determine the code for each noun sense that is dominated by that concept.

No.	Attributes				Class
	Outlook	Temperature	Humidity	Windy	
1	sunny	hot	high	false	N
2	sunny	hot	high	true	N
3	overcast	hot	high	false	P
4	rain	mild	high	false	P
5	rain	cool	normal	false	P
6	rain	cool	normal	true	N
7	overcast	cool	normal	true	P
8	sunny	mild	high	false	N
9	sunny	cool	normal	false	P
10	rain	mild	normal	false	P
11	sunny	mild	normal	true	P
12	overcast	mild	high	true	P
13	overcast	hot	normal	false	P
14	rain	mild	high	true	N

Figure 4.4: A data set to be modelled by a decision tree

are represented by a certain leaf belong to the same class (which is associated to that leaf). If the objects at a leaf node belong to different classes, then the class to which most of these objects belong is assigned to the leaf as a default class. Figure 4.5 shows a (perfect) decision tree which classifies the data in figure 4.4.

The usual goal of inducing a decision tree (in general, the usual goal of acquiring models in machine learning) is to make predictions about unseen data, i.e. to classify objects which are not included in the sample. For this purpose, it is not necessarily the best solution to learn a perfect decision tree. Such a tree would possibly overfit the data, e.g. capture noise in the training sample, which would result in a misclassification of unseen data. To find a balance between capturing regularities in the sample and avoiding the modelling of idiosyncracies, Quinlan & Rivest (1989) employ the MDL principle. To do this, the task of learning decision trees has to be reformulated as a problem of encoding the data sample by making use of a decision tree. This encoding is done in the following way:

(1) Encode the (possibly imperfect) decision tree

(2) Encode those objects that are misclassified by the decision tree (the list of exceptions)

The number of bits required for (1) is the model description length, the number of bits needed for (2) the data description length. The more complex the decision tree, the more bits are required to describe it. On the other hand, a more complex tree classifies more instances in the sample correctly so that less exceptions have to be encoded. According to the MDL principle, the decision tree which minimises the sum of the code length needed for (1) and (2) is selected.

A linguistic application of the MDL principle is described in (de Marcken 1996). The task is to

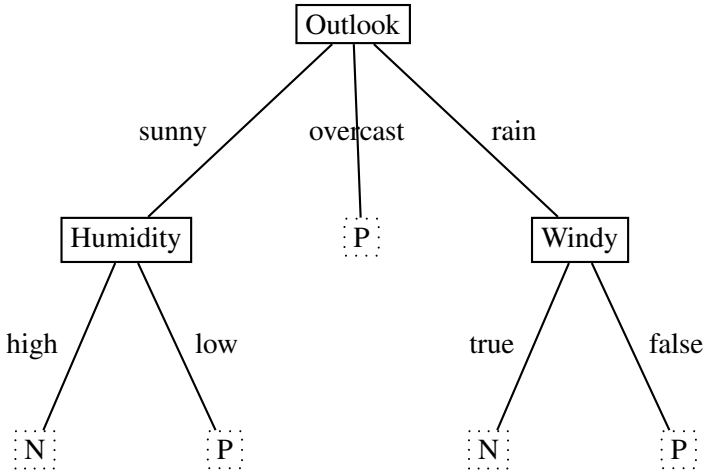


Figure 4.5: A decision tree

learn linguistic units from unsegmented text, i.e. text where word boundaries are not marked.[6] The simplest variant of the approach proposed by de Marcken aims at extracting words and multiword expressions from the examined corpus. The starting point is a lexicon whose entries represent single characters. During the learning process, new entries are created by concatenating existing entries. To decide which of the possible concatenations are added to the lexicon, de Marcken develops a learning method which is based on MDL. This method encodes the data sample (the unsegmented text data) by employing the lexicon. Each entry is assigned a unique binary code word. Following the rationale behind the design of coding schemes sketched in section 3.4.1.2, entries which have a high occurrence frequency in the sample are represented by short code words, while entries which rarely occur in the data receive long code words.

The learning algorithm iteratively adds new entries. In each iteration step, new candidate entries are created by the concatenation of existing entries. A new entry is represented in the lexicon as the concatenation of the code words of those entries whose concatenation created it. If such a candidate entry saves more bits for encoding the data than its representation requires, then this entry is added to the lexicon.[7]

Figure 4.6 and 4.7 show an artificial example that (de Marcken 1996, p. 41) provides to illustrate the idea behind his approach. Figure 4.6 shows three different lexicons which are capable of encoding the character sequence *themanonthemoon*. (A) is the most basic lexicon which is possible. It just contains entries for single characters. Note that the characters are ordered according to their frequencies in the data (*themanonthemoon* in this case): ‘o’, ‘n’ (3 occurrences); ‘t’, ‘h’, ‘e’, ‘m’ (2 occurrences); ‘a’ (1 occurrence). This ordering is crucial for assigning codewords. As can be seen, the codewords are ascendingly ordered by their lengths so that frequent entries are assigned shorter codes and and infrequent entries longer codes.[8] This code assignment analogously holds for the other lexicons. Lexicon

[6] This simulates one important condition of language acquisition, namely that a child perceives language as unsegmented phonetic signals.

[7] Actually, the algorithm is more complex. For example, useless entries may be deleted during the learning process.

[8] De Marcken applies Huffman coding, a widely used approach to code generation. Huffman coding generates a prefix-

(A)	Word	o	n	t	h	e	m	a	
	Codeword	00	01	100	101	110	1110	1111	
(B)	Word	o	n	the	m	t	h	e	a
	Codeword	00	01	100	101	1100	1101	1110	1111
(C)	Word	o	n	t	h	e	m	a	themanonthemoon
	Codeword	00	01	100	101	1100	1101	1110	1111

Figure 4.6: Alternative lexicons employed for encoding *themanonthemoon* (artificial example)

(A)	Evidence	100 101 110 1110 1111 01 00 01 100 101 110 1110 00 00 01 (t h e m a n o n t h e m o o n)
	Length	42 bits
(B)	Lexicon	1100 1101 1110 (t h e)
	Evidence	100 101 1111 01 00 01 100 101 00 00 01 (the m a n o n the m o o n)
	Length	40 bits
(C)	Lexicon	100 101 1100 1101 1110 01 00 01 100 101 1100 1101 00 00 01 (t h e m a n o n t h e m o o n)
	Evidence	1111 (themanonthemoon)
	Length	48 bits

Figure 4.7: Alternative encodings for *themanonthemoon*, employing the lexicons displayed in figure 4.6

(B) has an additional entry ‘the’, which is a concatenation of the entries ‘t’, ‘h’, and ‘e’. Likewise, lexicon (C) has an entry ‘themanonthemoon’, a concatenation of several character entries.

Figure 4.7 shows the description lengths which is needed to encode *themanonthemoon* by employing the different lexicons. One part of this description is the encoding of the respective lexicon, i.e. of the individual lexical entries. As all compound entries are (directly or via other compound entries) composed from the basic character entries, these entries have to be part of any lexicon. Thus, they do not have to be explicitly encoded. The compound entries are encoded by concatenating the codewords of their components. For example, the entry ‘the’ in lexicon (B) is represented by the concatenation of the codewords for ‘t’, ‘h’, and ‘e’.[9] The other part of the description is the coding of the evidence (data) by means of the codewords of lexicon entries. Since lexicon (A) only contains character entries, no lexicon entries have to be described. To describe the data, each character is encoded separately. Lexicon (B) has the entry ‘the’, which has to be encoded, but which on the other hand saves bits for encoding the evidence, since the portion *the* (occurring twice) can be described by one codeword

free code, i.e. no codeword is a prefix of any other codeword. This property is important to avoid ambiguities. As mentioned in section 3.4.1.2, the code generated by Huffman coding is guaranteed to be an optimal prefix code, i.e. yields the shortest code length on average. For details about Huffman coding, cf. (Cover & Thomas 1991, p. 92–101).

[9] Additionally, the order of the entries has to be encoded in order to determine the codeword for each entry. The example in figure 4.7 ignores this for the sake of simplicity.

instead of three with lexicon (A). Thus, the introduction of the parameter 'the' yields a shorter total description length. In contrast, the entry 'themanonthemoon' in lexicon (C) results in a higher overall description length. The encoding of the data can be significantly reduced to one codeword. However, the encoding of the entry itself is too costly.

As de Marcken reports, an experimental application of this approach created a lexicon whose entries to a large extent capture words and multiword expressions.

Above, I have argued that there is a correlation between the amount of data to be modelled and the complexity of the induced model which is inherent to the MDL framework: the more data have to be described, the higher the complexity of the acquired model. Of course, this also applies to both learning scenarios just sketched. More parameters (i.e. nodes in the decision tree or entries in the lexicon, respectively) become "affordable" if the amount of data to encode increases. However, in contrast to our tree cut models, this behaviour is perfectly appropriate for these applications. In the decision tree task, regularities induced from a large data set are less likely to reflect idiosyncracies than regularities found in a small data set. Thus, it is adequate to infer a more complex decision tree, which models more sophisticated regularities, if the training data set is large. The same holds for de Marcken's language acquisition task: New lexicon entries, which model a composition of existing entries, are only added if there is sufficient evidence for them (cf. the quotation on page 102). Overall, more data are likely to provide sufficient evidence for more lexicon entries.

To understand why the stated dependency between sample size and model complexity is appropriate for the approaches discussed here, but not for the Li and Abe approach, we have to look closer at the involved model classes. There is a crucial difference between decision trees and de Marcken's lexicons on the one hand and tree cut models on the other hand. In the former model classes, the simplest models (i.e. those models with the least parameters) are the most conservative models. They do not reflect any regularities in the data. The simplest decision tree is the empty tree. This is a trivial decision tree since it does not make any decisions, i.e. captures no regularities at all. For this tree, no parameter has to be encoded, but on the other hand, all data have to be coded as exceptions. The second simplest decision tree consists of one leaf node that is associated with one particular class.[10] This kind of tree merely models a default class. For such a tree, one parameter (the class assigned to the node) has to be encoded; all data items which are classified differently have to be encoded as exceptions. The more sophisticated a decision tree, the more subtle regularities it models, and hence, the more subtle predictions it makes about unseen data. In de Marcken's approach, the simplest lexicon (the starting point of the learning algorithm) just contains entries of the letters of the assumed alphabet. This lexicon does not make any predictions about which combinations of these letters form words or multiword expressions. The more new parameters (which make exactly that kind of predictions) are introduced in the learning process, the stronger hypotheses regarding the probable compositions of letters and words are captured by the resulting model. Summing up, the two model classes have in common that complex models express stronger claims about the regularities in the data, while simple models are more conservative. It is important to note that this applies to many well-known model classes which have been proposed in the literature in connection with the MDL paradigm, be it for linguistic (e.g. Osborne's (1997) induction of categorial grammars) or non-linguistic tasks (e.g. the task of acquiring a polynomial function from a set of points in a coordinate system, cf. (Li & Vitányi 1992), (Rissanen & Ristad 1992)).

Actually, the class of tree cut models representing selectional preferences differs from that pattern. The simplest tree cut model is a cut located at the root of the hierarchy. This model comprises only

[10]The learning algorithm in (Quinlan & Rivest 1989) starts from trees of that kind.

one parameter (the preference value of the root concept). However, it is not the most conservative tree cut model. In fact, it is the least conservative model. It generalises from the training data as much as possible by representing all noun complements of the verb under consideration with the same preference value. Thus, such a model states that the examined verb does not exhibit selectional preferences or dispreferences at all. This is the strongest claim that is conceivable, since it completely denies the phenomenon of selectional preference, at least for a particular verb. For the task of inducing selectional preferences from a data sample, a conservative model should represent selectional preferences in a way that reflects the information provided by the data, but no further inferred regularities. In the tree cut model class, regularities are expressed by generalisations. Therefore, the most conservative tree cut is located at the leaves of the hierarchy so that each noun sense in the sample is represented by a separate parameter,[11] and the probabilities determined by the model correspond to the MLE probabilities obtained from the sample. But following the MDL paradigm, the tree cut model at the leaves is the most complex model, since it contains the most parameters.

Hence, the desired dependency between the amount of available training data and the amount of inference expressed by the acquired model, which holds in many model classes, is reversed in the tree cut model class: Strictly applying the MDL principle here, more data (and thus more evidence) yield a more conservative model, whereas less data yield a model which makes stronger claims about the distribution underlying the data. Of course, this is inappropriate, since it is the opposite of the desired correlation that more evidence should trigger stronger inferences. By *strictly* applying MDL, I mean selecting that model which minimises the "plain" sum of model and data description length. In the following section, I will propose a variant of this approach where the quantity to be minimised is a slight modification of that sum. With this modification, the adequate coincidence between sample size and generalisation of the model is established.

4.2 Balancing Model and Data Description Length

4.2.1 Introducing a Weighting Factor

Recall that the Li and Abe approach selects the tree cut model which minimises the sum of the parameter description length L_{par} (the number of bits required to encode the preference values of the concepts on the cut) and the data description length L_{dat} (the number of bits needed to encode the data sample employing the model). The sample S_v to be encoded consists of the noun complements of the examined verb v. Technically, the problem discussed in the previous section is caused by different complexities of L_{par} and L_{dat} with respect to the sample size $|S_v|$. Both complexities depend on $|S_v|$, but in a different manner. As one can see from the equations (3.31) (or (4.1), respectively) and (3.32) on page 78, L_{par} has the complexity $O(\log |S_v|)$ (each parameter is represented by a precision of $\frac{\log |S_v|}{2}$ bits), while L_{dat} has the complexity $O(|S_v|)$ ($|S_v|$ data items have to be encoded). Thus, with growing sample size, L_{dat} "grows faster" than L_{par}, and for frequent verbs, the model description length can be neglected, so that a model with many specific concepts becomes "affordable".

To overcome this drawback, I modified the learning algorithm (step 9. in figure 3.16 on page 85) in the following way: I extended the expression which is to be minimised by a weighting factor. Instead of minimising $L_{par} + L_{dat}$, the modified algorithm minimises

[11] Note that according to equation (4.1) on page 99, the model only contains parameters for noun senses which occur in the sample.

$$L_{par}(M) + C\left(\frac{\log|S_v|}{|S_v|}\right) L_{dat}(M) \quad (C > 0) \tag{4.2}$$

In this formula, L_{dat} is weighted by the factor $C \times \frac{\log|S_v|}{|S_v|}$. This quantity itself comprises two parts. The fraction $\frac{\log|S_v|}{|S_v|}$ reduces the complexity of L_{dat} from $O(|S_v|)$ to $O(\log|S_v|)$. Thus, both addends have the same complexity. $|S_v|$ does not directly affect the level of generalisation any more.[12] Instead, the content of the sample, i.e. its "semantic variety", is crucial for triggering generalisation (cf. the discussion at the beginning of section 4.1.2). This is the desired behaviour.

C, the other part of the weighting factor, is a positive constant which may be arbitrarily chosen. The value of C affects the proportion of the data description length to the total quantity subject to minimisation. A high value of C biases the learning procedure towards acquiring a rather complex model, because the code savings which such a model provides are weighted higher than its costs. Conversely, a low value of C introduces a bias towards acquiring a rather simple model, since the costs of encoding a model with more parameters get a higher weight than the savings such a model brings about. In other words, C influences the degree of generalisation: The smaller C is, the more general concepts are acquired. I introduced this constant to account for the discussion in section 3.1: The possibility of manipulating the overall generalisation level by the choice of C introduces some flexibility which might prove useful to fine-tune the approach w.r.t. a specific application scenario (i.e. a specific task, domain, language, etc.). An appropriate value of C should be found empirically for a particular application. Note that C is intended to be independent of a particular verb. It provides a global bias regarding the amount of generalisation. Relative to this bias, the content of the sample belonging to a particular verb triggers generalisation specifically for that verb.

4.2.2 Theoretical Justification

The introduction of a weighting factor is a deviation from the "pure" MDL principle which states that the model minimising the total description length is the optimal one. However, one can find usages of the term *MDL principle* in the literature (e.g. (Quinlan & Rivest 1989)) which mean something like "finding an appropriate balance between model simplicity and model accuracy w.r.t. the training data". In this sense, the weighting approach which I propose is still compatible with the MDL framework. In any case, it can be shown that this approach is compliant to Bayesian reasoning. I will proof this claim in this subsection.

Bayesian reasoning (*Bayesian inference*, *Bayesian learning*) is a learning paradigm which is based on Bayes' law:

$$p(M|S) = \frac{p(M)p(S|M)}{p(S)} \tag{4.3}$$

If we interpret S as referring to a data sample (the available evidence) and M as referring to a model, then $p(M|S)$ (the *posterior probability*) is the crucial distribution we are interested in, since it guides

[12] However, as discussed in section 4.1.2, an indirect correlation between sample size and generalisation is apparent: Frequent verbs tend to have more general preferences. This tendency evolves from distributional properties of language rather than mathematical properties of the employed formula. It will be empirically supported by the experiments reported later in this chapter.

the selection of a model M given the data S: of course, one selects the model that maximises $p(M|S)$. Bayes' law (which immediately follows from elementary probability theory) states that this probability can be obtained by the product of the *prior probability* $p(M)$ (the model probability without any knowledge about evidence) and the *likelihood probability* $p(S|M)$, divided by $p(S)$, the probability of the data sample regardless of a particular model.

The weighting approach can be reformulated in terms of Bayesian reasoning.[13]. To achieve this, the crucial step is to appropriately convert (weighted) code lengths to probabilities. The basic idea of this conversion is adopted from (Quinlan & Rivest 1989). They provide a formula for assigning a probability to an arbitrary binary string b with length $L(b)$:

$$p(b) = \left(1 - \frac{1}{r}\right)\left(\frac{1}{2r}\right)^{L(b)} \tag{4.4}$$

$r > 1$ is an arbitrary, but fixed constant. This formula defines a probability distribution over binary strings. As there exist 2^L binary strings of length L, all strings of length L together have the probability $\left(1 - \frac{1}{r}\right)\left(\frac{1}{r}\right)^L$. For $r > 1$, the sum over all lengths L of $\left(\frac{1}{r}\right)^L$ converges to $\frac{1}{1-1/r}$ (this is the well-known geometric series). Multiplying by $\left(1 - \frac{1}{r}\right)$ yields 1. Thus, equation (4.4) defines a proper probability distribution. The higher the value of r is chosen, the faster the string probabilities decrease with the string lengths.

To convert coding schemes (rather than arbitrary binary strings) to probability distributions, equation (4.4) presented in (Quinlan & Rivest 1989) requires a slight modification. Given a particular coding scheme, there might be binary strings which are not used by that coding scheme. In particular, this applies for prefix codes, where strings which form a prefix of any employed code word are not used themselves. Thus, if code words were assigned probabilities according to equation (4.4), then the probabilities of all possible code words would not sum to 1. Thus, the formula has to be modified by replacing the normalising constant $\left(1 - \frac{1}{r}\right)$ by a general constant N whose value depends on the actual coding scheme:

$$p(c) = N\left(\frac{1}{2r}\right)^{L(c)} \tag{4.5}$$

where

$$N = \frac{1}{\sum_{c\text{: } c \text{ is possible code word}} \left(\frac{1}{2r}\right)^{L(c)}}$$

It will turn out that for our purposes, it is not necessary to know the exact value of N, because this constant does not play a role for the selection of the model M. We know that N exists and is limited by $1 - \frac{1}{r}$.

[13] Likewise, the "original" tree cut approach can be reformulated as a Bayesian learning problem (cf. (Li & Abe 1998)). Moreover, the MDL principle as such can be formalised in Bayesian terms (cf. (Mitchell 1997), (Li & Vitányi 1992), (Osborne 1997)).

Employing probability distributions as defined in equation (4.5), it can be shown that the minimisation of expression (4.2) has an equivalent Bayesian formalisation. First, define the prior probability of model M as

$$p(M) = N_{mod} \left(\frac{1}{2r_{mod}} \right)^{L_{mod}(M)} \tag{4.6}$$

and the (likelihood) probability of sample S given M as

$$p(S|M) = N_{dat} \left(\frac{1}{2r_{dat}} \right)^{L_{dat}(M)} \tag{4.7}$$

As noted above, Bayesian inference selects that model that maximises the posterior probability $p(M|S)$. The maximisation of $p(M|S)$ is equivalent to the minimisation of $-\log p(M|S)$. Furthermore,

$$\begin{aligned}
& \arg\max_M \quad p(M|S) \\
= & \arg\min_M \quad -\log p(M|S) \\
= & \arg\min_M \quad -\log \frac{p(M)p(S|M)}{p(S)} \\
= & \arg\min_M \quad (-\log p(M) - \log p(S|M) + \log p(S)) \\
= & \arg\min_M \quad (-\log N_{mod} - L_{mod}(M) \log(1/2r_{mod}) \\
& \qquad\quad -\log N_{dat} - L_{dat}(M) \log(1/2r_{dat}) \\
& \qquad\quad +\log p(S)) \\
= & \arg\min_M \quad (L_{mod}(M)(1+\log r_{mod}) + L_{dat}(M)(1+\log r_{dat}) \\
& \qquad\quad -\log N_{mod} - \log N_{dat} + \log p(S)) \\
= & \arg\min_M \quad (L_{mod}(M)(1+\log r_{mod}) + L_{dat}(M)(1+\log r_{dat}))
\end{aligned}$$

In the last line, all addends that do not depend on M are omitted, since they do not play a role for the minimisation.

Let D be a constant so that

$$D > \max\left\{ \frac{1}{|S|}, \frac{1}{C \log |S|} \right\} \qquad (C > 0)$$

Now we set

$$\begin{aligned}
r_{mod} &= 2^{D|S|-1} \\
r_{dat} &= 2^{DC \log |S|-1}
\end{aligned}$$

The condition on D ensures that r_{mod} and r_{dat} are greater than 1. Therefore,

$$\begin{aligned}
&\arg\min_M \quad (L_{mod}(M)(1+\log r_{mod}) + L_{dat}(M)(1+\log r_{dat})) \\
=&\arg\min_M \quad (D|S|L_{mod}(M) + DC\log|S|L_{dat}(M)) \\
=&\arg\min_M \quad \left(L_{mod}(M) + \frac{C\log|S|}{|S|}L_{dat}(M)\right)
\end{aligned}$$

As $L_{mod}(M) = L_{cut}(M) + L_{par}(M)$ and $L_{cut}(M)$ is constant for all models (cf. section 3.4.3.2), the last line is equivalent to minimising (4.2).

Intuitively, $p(M)$ and $p(S|M)$ are defined by equations (4.6) and (4.7), respectively, in a way that their decrease with growing L_{mod} and L_{dat} is comparable. Thus, none of these two probabilities will overly "dominate" the other one.

At the end of this subsection, I would like to briefly address the criticism of the Bayesian framework put forward by Rissanen & Ristad (1992). They claim that the prior probability distribution may be arbitrarily designed, while the design of its MDL counterpart, the model description, is guided by the "universal" criterion of efficient encoding:

> ... in the Bayesian approaches the prior can in principle be selected as we wish. In fact, it often represents private subjective knowledge. ... However, the MDL framework enjoys significant conceptual and practical advantages over the Bayesian framework. It is often easier to design a prefix code than to design a probability function directly. Conversely, it it is often more difficult to to design an unnatural prefix code than to design an arbitrarily unnatural probability function. (Rissanen & Ristad 1992, p. 160,164)

Let me emphasise that this criticism does not apply to the weighting approach which I have introduced here. The model prior $p(M)$ is not selected arbitrarily, but is determined by the complexity of the tree cut model M and a well-motivated coding scheme. The only "subjective" component of $p(M)$ (due to its relation to $p(S|M)$) is the choice of the constant C. However, this selection is intended to be guided empirically, by evaluating the approach in a particular application, trying different values of C and selecting that value that yields maximal performance.

4.2.3 Experimental Results

To test the impact of weighting on the generalisation level of the acquired tree cuts, I examined verbs with diverse numbers of *different* noun complements (types) in the training sample mentioned in section 4.1.1. In particular, I selected all verbs with a high number (≥ 1000), a medium number (400–600), a low number (70–100), and a very low number (10–40) of different complements and compared the generalisation level retrieved by the "standard MDL" algorithm and the weighting algorithm. (I arbitrarily chose $C = 50$.) For all verbs with a high number and 89% of the verbs with a medium number of different complements, the weighting algorithm obtained more general concepts than the standard MDL algorithm. In contrast, more specific concepts were computed for almost all verbs with a low and a very low number of different complements (95.9% and 99.5%, respectively). Hence,

$ncpt$	$A(kill, ncpt)$
...	...
<life_form#organism#being#living_thing>	3.19
...	...

Figure 4.8: Part of the tree cut model for “kill” (weighting algorithm)

$ncpt$	$A(murder, ncpt)$
...	...
<person#individual#someone#mortal#human>	4.32
<animal#animate_being#beast#brute>	0.27
<plant#flora#plant_life>	0.08
...	...

Figure 4.9: Part of the tree cut model for “murder” (weighting algorithm)

the modification changes the behaviour of the algorithm towards the desired direction: variety of complements triggers generalisation.

Figures 4.8–4.10 show the tree cut models for “kill”, “murder”, and“assassinate” which are yielded by the weighting algorithm ($C = 50$). Now these models exhibit an appropriate level of generalisation. (Note that WordNet does not have a concept like “important person”, which would be appropriate for “assassinate”. However, the concepts displayed in figure 4.10 fall under this category without exception.)

As finding the appropriate generalisation level is crucial for learning appropriate thematic role relations, the preliminary experiments described in section 4.1.1 and in this section strongly indicate that introducing a weighting factor as described in section 4.2.1 significantly improves the suitability of the tree cut approach for the task of this thesis. The next chapter describes design and preprocessing considerations of experiments to verify that claim, as well as their results. These results show that indeed the weighting approach drastically outperforms the standard MDL approach.

$ncpt$	$A(assassinate, ncpt)$
...	...
<adult>	2.57
<capitalist>	1.07
<communicator>	5.22
<contestant>	8.62
<disputant#controversialist>	5.02
<spiritual_leader>	23.10
<head#chief#top_dog>	71.55
<head_of_state#chief_of_state>	242.15
<presiding_officer>	356.61
<REST::leader>	653.89
<peer#equal#match#compeer>	18.53
<relative#relation>	7.12
<party>	68.31
<ruler>	55.49
<authority>	197.08
<suspect>	653.79
...	...

Figure 4.10: Part of the tree cut model for “assassinate” (weighting algorithm)

Chapter 5

Acquiring Selectional Preferences for Thematic Role Relations: Practical Issues

While the previous chapter has dealt with the core approach I propose for learning selectional preferences, this chapter addresses more practical, in a sense peripheral, but nonetheless important issues. These issues concern preprocessing steps on the one hand and the evaluation of the approach (and related approaches) on the other hand.

Section 5.1 describes the data employed for training and discusses their usage. Section 5.2 and 5.3 describe extensions and refinements of the preprocessing step of mapping word frequencies to concept frequencies. This task is divided into two steps: lexical disambiguation of the training data and propagating frequency counts of word senses to more abstract concepts in the hierarchy.

Section 5.4 deals with the issue of retrieving a gold standard for thematic role relations from the EuroWordNet database. With this gold standard, it is possible to perform a systematic and application-independent evaluation of approaches for learning selectional preferences. While the experiments referred to in section 4.1.1 and 4.2.3 only provided illustrative examples, section 5.5 presents more recent experiments which employ the gold standard and thus operate on a broader empirical basis. These experiments show that the weighting approach which I propose significantly outperforms the original approach developed by Li and Abe.

5.1 The Training Data

This section describes the training data I used for the experiments which I report in this chapter and the chapters 6 and 7. These data are different from the data I used for the preliminary experiments sketched in the previous chapter. Both have been retrieved by parsing the British National Corpus (cf. (Burnard 1995)). However, the data I referred to in section 4.1.1.2 only capture part of the BNC (about 40 million words), whereas the data I describe here cover the whole BNC (100 million words). More importantly, the former data were obtained by employing a chunk parser, while the latter data have been extracted from complete parse trees. A chunk parser is fairly suitable for determining the subject or the object of an English sentence. However, to acquire the complete subcategorisation frame of a sentence, a complete parse is required. In particular, a chunk parser does not account for PP-attachment disambiguation, i.e. for determining whether an occurring PP is a complement of a

verb or modifies a noun. This kind of information is needed to detect thematic roles which underlie the syntactic realisation of the arguments in a sentence. For example, role relations involving a location, i.e. LOCATION, DIRECTION, SOURCE_DIRECTION, and TARGET_DIRECTION, are typically realised as PPs. Moreover, as we saw in section 2.1 (cf. also the discussion of the examples (5.1)–(5.3) below), the role which underlies a certain syntactic complement can be partially identified by the kinds of syntactic complements co-occurring in the sentence. For these quantitative and qualitative reasons, the data I will describe in this section are more adequate for the task of this thesis than the data used in the experiments sketched in chapter 4. Since these experiments were restricted to direct objects, the shallow-parsed data were sufficient. Furthermore, the data to be introduced here had not been available when these early experiments were carried out.

The training data I used had been compiled by Sabine Schulte im Walde (cf. (Schulte im Walde 1998*a*, p. 11–31) for further details of retrieving and preparing the data) at the IMS, University of Stuttgart.[1] These data were retrieved by parsing the BNC with a stochastic parser which had been developed at the IMS (cf. (Carroll & Rooth 1998)). For each parsed sentence, the head verb, the subcategorisation frame representing the sentence's syntactic structure, and the head word of each complement were extracted. All extracted word forms were lemmatised by employing a morphological lexicon and a morphological stemmer. Frames of passive sentences were transformed into the corresponding active frames (the subject became the direct object and the by-phrase, if present, became the subject). Sentences headed by an auxiliary verb were excluded. This procedure yielded a set of 3,428,273 (representations of) sentences. An example of such an item is

```
give subj:saleswoman obj:apple pp.to:customer
```

As explained in section 3.2, the approaches investigated in this thesis do not take into account dependencies between the different complements within a sentence. Thus, they do not take as input a verb's complete collection of complements in a sentence, but rather process simple verb–complement pairs. Therefore, to obtain the data in an appropriate format, the complete frame items obtained as described above were divided into items each of which represents the relation between a verb and a single complement as well as additional syntactic information. For example, the frame item above corresponds to the following data items:[2]

```
give#subj:obj:pp.to/subj  saleswoman
give#subj:obj:pp.to/obj  apple
give#subj:obj:pp.to/pp.to  customer
```

The left component of these pairs consists of three parts.The first part is the verb. The second part (separated from the verb by a '#'[3]) represents the syntactic arguments the verb takes in the sentence from which the data item originated, i.e. the complete subcategorisation frame realised in that sentence. The individual arguments are separated by colons. NP arguments are characterised by their grammatical function (subject or object), PP arguments by the "pp" label and their preposition ("to"

[1] I am grateful to Sabine Schulte im Walde for making these data available to me.

[2] In the files I received from Sabine Schulte im Walde, the compiled collection of data items is represented in a much more compact form; cf. (Schulte im Walde 1998*a*) for details.

[3] The notation illustrated here is similar to the notation used by Schulte im Walde. In addition, I had to introduce the separators '#' and '/' for implementational reasons.

in the examples above). Of course, other argument types (e.g. clausal complements) are represented as well. However, since they play a marginal role for my work, I will refer to them only where necessary. The third part (separated by a '/') indicates that syntactic argument type, i.e. that syntactic relation, that the data item actually represents.[4] Of course, this argument type belongs to the subcategorisation frame represented in the second part. The right component of a data item contains the head noun of that argument. For example, the item

```
give#subj:obj:pp.to/obj  apple
```

represents a corpus instance where the verb "give" subcategorises for a subject, an object, and a prepositional phrase headed by the preposition "to", and the object is "apple".

The additional syntactic information provided with the verb is necessary to select those data items that correspond to that argument for which selectional preferences shall be acquired. It is obvious that the third part of the left component is required if one learns preferences for a certain syntactic argument of a verb, as has been done in all the work discussed so far. For example, to induce selectional preferences for the subject, one selects those tuples which are marked with `.../subj` for training. The information about the complete syntactic subcategorisation pattern (in the second part of the left component) is important if selectional preferences for a semantic argument (i.e. a thematic role), rather than a syntactic argument, are to be learned. As noted above, the semantic role underlying a syntactic argument can to a certain extent be identified (or, at least, the range of possibilities can be narrowed) by taking into account the other syntactic argument types in the sentence. Thus, the information about the syntactic subcategorisation pattern is important for linking. Obviously, a certain linking procedure, i.e. a mechanism to map syntactic to semantic arguments, is necessary to employ syntactic training data for acquiring thematic role relations. In our acquisition context, a more concrete definition of the linking task is the following: Given a certain thematic relation type to be learned, select all items in the data which represent a realisation of that thematic relation. These items are used tor train the acquisition algorithm so that this algorithm learns selectional preferences for the role in question.

To illustrate such a linking process, recall examples (2.2)–(2.4), repeated here, slightly modified, as (5.1)–(5.3):

(5.1) The jealous husband broke the window.

(5.2) A hammer broke the window.

(5.3) The jealous husband broke the window with a hammer.

A syntactic alternation of the kind illustrated by these examples is called *diathesis alternation*.[5] Such alternations have been largely examined for English verbs (cf. (Levin 1993)). This research is guided

[4] In figure 3.4 on page 61, I used a different notation to express the sort of information which is provided by the left component of a data item. For instance, the expression "increase (subj obj)" in that figure corresponds to `increase#subj:obj/obj` in the notation just described. Although the former notation might appear more concise, I will stick to the latter one for the rest of this thesis, because it makes explicit the important fact that the left component comprises three different pieces of information (a verb, a subcategorisation frame, and a particular syntactic relation).

[5] As we will see in section 6.3, there are a number of types of diathesis alternation. The kind of alternation exemplified in (5.1)–(5.3) is called *instrument subject alternation*.

by the hypothesis that semantic properties of verbs are reflected by their syntactic behaviour, and that a semantic classification of verbs can be induced from the types of diathesis alternation that each verb does or does not undergo. Note that Schulte im Walde created and used her training data exactly for the task of automatically acquiring such a verb classification. I will utilise linguistic findings about diathesis alternations in my linking strategy (cf. section 6.3).

In the sentences above, "husband" is the Agent, "window" the Patient, and "hammer" the Instrument. These examples illustrate that linking is all but trivial: Both Agent and Instrument can be realised as the subject. Furthermore, the Instrument can be realised as the subject or a with-PP. However, it is not possible that both the subject and a with-PP express an Instrument in the same sentence. Finally, neither the Agent nor the Instrument is expressed as the object, since the object expresses the Patient. Such dependencies between arguments can be employed in the linking process.

Now suppose that the sentences above belong to the training corpus. Then, if one wants to acquire selectional preferences for the INSTRUMENT role, the following items should be included in the training set, where the first item corresponds to sentence (5.2), while the second one corresponds to (5.3):

```
break#subj:obj/subj  hammer
break#subj:obj:pp.with/pp.with  hammer
```

However, to learn preferences for the AGENT role, the following items are relevant (the first item corresponding to (5.1), the second one to (5.3)):

```
break#subj:obj/subj  husband
break#subj:obj:pp.with/subj  husband
```

As this is a very sophisticated task (which, indeed, is worth a Ph.D. thesis of its own), I will postpone this issue here and dedicate a separate chapter to it (chapter 6).

To eliminate noise, which could arise due to parsing errors or idiosyncratic syntactic usages of verbs in the corpus, I only took those data items into account whose left component (verb + frame + argument) occurs at least 10 times in the data and whose frame-argument combination appears in at least 5% of all occurrences of the respective verb in the data.

5.2 Disambiguating the Training Data

This and the next section deal with the issue of mapping the word forms in the data to concepts in the WordNet hierarchy. This preprocessing step is necessary because our goal is to acquire relations between semantic verb and noun *concepts*, rather than verb and noun *forms*. Following the basic idea of the word-to-sense approach, this problem comprises two sub-tasks: the semantic disambiguation of word forms (i.e. mapping forms to senses) and passing the evidence found for individual word senses to those concepts which subsume these senses.[6] This section addresses the former of these sub-tasks,

[6]The word-to-concept approach does not distinguish these two steps, since it immediately projects the frequency of a word form to all the concepts that subsume that form, i.e. it treats the senses corresponding to that form and their hyperonyms in a uniform way.

while the next section is concerned with the latter.

Unlike Li and Abe, who acquire the selectional association between verb forms and noun concepts, I aim at learning the selectional association between verb concepts and noun concepts. To achieve this, the verbs in the training data have to be disambiguated as well as the nouns. As the training data consist of a collection of verb–noun pairs[7], the task of disambiguating these data can be formalised in the following way: Given a verb–noun pair (v, n), we have to estimate the probability $p(sns_i(v), sns_j(n)|v, n)$ (where $sns_i(w)$ denotes sense i of word w) for each possible combination $(sns_i(v), sns_j(n))$ of senses of v and n. This information can be employed to estimate the frequency counts of verb–noun *sense pairs* $freq(sns_i(v), sns_j(n))$ from the counts of verb–noun *form pairs* $freq(v, n)$ extracted from the training data:

$$freq(sns_i(v), sns_j(n)) = freq(v, n) \times p(sns_i(v), sns_j(n)|v, n) \tag{5.4}$$

As in chapter 3, I leave the syntactic information (provided in the individual items of the training data as the second and third part of the left component, s.a.) implicit. Frequencies and probabilities concerning a verb–noun pair are computed separately for each syntactic configuration associated with that pair in the data.

Of course, words which are not covered by WordNet cannot be taken into account; they are dropped in the step of mapping word forms to word senses. Proper nouns are a special case. Some researchers employ strategies for a coarse semantic classification of proper nouns, in order to map them to very general WordNet concepts. For example, McCarthy (2001) uses named entity recognition software to assign proper noun instances to one of the concepts <person#someone>, <organization>, and <location>. In principle, such a strategy makes sense. As usually a considerable amount of verb complements are proper nouns, capturing these complements significantly increases the data size, and thus the empirical evidence actually employed for learning selectional preferences. On the other hand, since such techniques provide only a very general semantic characterisation of proper nouns, they might introduce a significant bias to very general WordNet concepts, especially if proper nouns constitute a major portion of the data. This bias would thwart the process of generalising from the data instances to find the appropriate level of abstraction (cf. the discussion at the beginning of section 4.1.2). The problem is that the coarse classification of named entities involves a high amount of abstraction, and thus the acquisition algorithm would have to generalise from very abstract items (while proper nouns themselves by definition are maximally specific!). For example, recall the selectional preferences for the object of "assassinate" discussed in the previous chapter. The appropriate generalisation level for these preferences is *below* <person>, because this verb is used only in connection with a specific type of persons (important persons). The tree cut acquired for "assassinate" (cf. figure 4.10 on page 114) adequately models this generalisation level. If, however, a large portion of the objects of "assassinate" are proper nouns, and if all these nouns are represented by <person>, then the collection of noun senses from which the tree cut acquisition algorithm has to generalise includes <person> with a considerable frequency. Under these circumstances, it would be likely that the learned tree cut model contains the concept <person> with a rather high preference value, which would be an over-generalisation. Therefore, I decided not to give a special treatment to proper nouns. In other words, proper nouns which do not occur in WordNet (i.e. most proper nouns) are not taken

[7]For the sake of simplicity, the additional syntactic information provided by the data items as described in the previous section is not taken into account for word sense disambiguation here. In principle, the knowledge of the actual subcategorisation frame might exclude some of the possible senses of a verb. Employing this kind of knowledge would be a refinement of the WSD method to be proposed which is worth investigating in future work.

into account for learning selectional preferences.

5.2.1 The Uniformity Hypothesis

Lacking further information concerning sense distributions of word forms, it is reasonable to make the following simplifying assumptions, which I will henceforth refer to as the *uniformity hypothesis*:

(a) a uniform distribution of the senses of v and n, respectively

(b) independence of the sense distributions of v and n

With these assumptions, the probability of senses given word forms can be estimated as in equation (5.5):

$$p(sns_i(v), sns_j(n)|v, n) = \frac{1}{senses(v)} \times \frac{1}{senses(n)} \tag{5.5}$$

(For example, if v has 9 and n 7 possible senses, then each combination of these senses receives the probability $\frac{1}{9} \times \frac{1}{7} = 0.016$.) Almost all the strategies for learning selectional preferences discussed so far rest on the uniformity hypothesis.[8] Only Agirre & Martinez (2002) and in part Ribas (1995*a*) circumvent the problem by training on lexically disambiguated data. Indeed, if no disambiguation information is available, there does not seem to be any obvious alternative to assuming uniformity. Apart from that "negative" justification, Resnik provides a positive motivation for the appropriateness of the uniformity hypothesis. As he puts forward his argument in several publications concerning the acquisitions of selectional preferences (cf. (Resnik 1993), (Resnik 1997), (Resnik 1998)), and as I think that this argument is not unproblematic, I will address it here. He argues as follows:

> Note that, in the absence of sense disambiguation, it is not unreasonable to distribute the credit for an observed noun equally among all the classes subsuming it [this realises the uniformity hypothesis within the word-to-concept approach, AW], as a first approximation. This works because related words tend to be ambiguous in different ways. For example, consider the observation of two verb–object combinations, *drink wine* and *drink water*. On the basis of these observations, the joint frequency will be incremented for each class containing *wine* in any sense—including, for example, <chromatic_color>. Similarly, the second pair will be recorded as a co-occurrence between *drink* and inappropriate categorisations of *water* such as <body_of_water>. However, evidence for co-occurrence will *accumulate* only for classes containing both *water* and *wine*, such as <beverage>. The cumulative evidence of co-occurrence with *drink* will thus tend to support appropriate interpretations, and counts with inappropriate senses will appear only as low frequencies dispersed throughout the taxonomy. (Resnik 1998, p. 251)

The critical part of this argument is the claim that "related words tend to be ambiguous in different ways". In fact, more recent WordNet-related work has pointed out that this is not necessarily the case.

[8] For those approaches that acquire preferences of verb forms, only assumption (a) (applied to nouns) is relevant.

This work deals with enriching WordNet with information concerning a well-known phenomenon in lexical semantics: *systematic polysemy* (*regular polysemy*). This term refers to the well-known fact that different words have analogous patterns of senses. A widely discussed example of such a pattern is the collection of senses exhibited by the word "school" (cf. (Bierwisch 1983)):

(5.6)

a. The school went for an outing.

b. School starts at 8.30.

c. The school was founded in 1910.

d. The school has a new roof.

These examples (cited from (Buitelaar 1998)) demonstrate different senses of "school": In (5.6 a.), this word means a group of people, in (5.6 b.) a process, in (5.6 c.) an institution, and in (5.6 d.) a building. This pattern of senses is regular, i.e. it is (completely or in parts) shared by a number of other words such as "university", "kindergarten", "theatre", "opera", "parliament", etc. Other systematic sense patterns mentioned in the literature include *container / containerful* ("glass", "cup"), *animal / food* ("lamb", "chicken"), *language / people* ("Spanish", "German"); cf. (Ostler & Atkins 1992) for further examples. In contrast, *homonymy* refers to non-systematic sense patterns that are idiosyncratic for a particular word. Usually, this phenomenon is exemplified in the literature by the word "bank", which has the two completely unrelated senses 'waterside' and 'financial institution', illustrated in (5.7) (again cited from (Buitelaar 1998)):

(5.7)

a. We walked along the bank of the Charles river.

b. Did he have an account at the HBU bank?

In our context, the crucial question is to what extent the phenomenon of systematic polysemy is reflected by the sense distributions of the words captured by WordNet. Recently, several approaches to detect patterns of systematic polysemy in WordNet have been proposed (cf. (Buitelaar 1998), (Buitelaar 2000), (Peters & Peters 2000)). I will not go into details of these approaches here. They share the common idea to group words together whose sense patterns can be characterised by a common set of superconcepts. For example, Peters & Peters (2000) found a group of 20 words which can be characterised by either of the two concepts <music> and <dance> ("waltz", "rumba", "bolero" etc.). Each of these words has (at least) one sense in WordNet that is subsumed by <dance> and one sense subsumed by <music>. Hence, this group comprises instances of the systematic polysemic pattern *music / dance*. Other patterns Peters & Peters (2000) found in WordNet are e.g. *musical composition / group of singers* (16 words, e.g. "trio", "quartet", "suite"), *building / institution* (15 words, e.g. "school", "chamber", "court"), *supporting structure / theory* (8 words, e.g. "framework", "foundation", "base") and *container / quantity* (33 words, e.g. "barrel", "firkin", "kettle").

Moreover, Buitelaar (1998) found sense patterns that are *coincidentally* shared by several words. For example, the pattern *act / animal / artifact* is shared by the words "bat", "drill", "fly", "hobby", "ruff", "solitaire", and "spat". Other patterns which reflect systematic relationships between senses, like *act / region* (e.g. "caliphate", "emirate", "clearing", "repair") may refer to an act or to a location where that

act takes or took place), also have coincidental instances in WordNet (e.g. "bolivia" or "chicago" may refer to a location or a card game, "charleston" to a location or a dance). Buitelaar (1998) summarises his findings by claiming that almost 95% of the nouns captured by WordNet have senses that are "somehow related", and only about 5% "are to be viewed as true homonyms", exhibiting completely unrelated senses.

These results contradict Resnik's assumption that "related words tend to be ambiguous in different ways". Hence, his conclusion that "counts with inappropriate senses will appear only as low frequencies dispersed throughout the taxonomy" is questionable. Rather, if related words exhibit common sense patterns, then there is the danger that the frequency counts of erroneous senses accumulate as well as the counts of the correct senses. Ribas (1994) reports that the accumulation of erroneous senses resulted in the acquisition of inappropriate selectional preferences. My experimental evidence confirms this observation. Regarding the findings sketched above, this seems an obvious consequence. It remains an empirical question to what extent this is a problem for learning selectional preferences. Therefore, I carried out the experiments reported below in two ways: once adopting the uniformity hypothesis, and once employing a strategy of lexical disambiguation which I describe in the following section.[9]

5.2.2 A Disambiguation Approach

The previous section shows that it makes sense to obtain a more informed estimation of sense distributions than to rely on the uniformity hypothesis. Therefore, I used a more sophisticated approach to estimate $p(sns_i(v), sns_j(n)|v, n)$. This approach consists of two steps: The first step employs the latent semantic clustering technique (LSC) developed by Rooth et al. (1998) and introduced in section 3.3 to obtain semantic clusters of similar verb–noun pairs. The second step employs a method proposed in (Resnik 1995*a*) which disambiguates words within a semantic cluster by measuring the semantic distances between their respective senses.

In the first step, I acquired latent semantic classes from the training data described in section 5.1. Like in (Rooth et al. 1998), subcategorisation information was attached to the verbs. To reduce the number of parameters (class membership probabilities) to be estimated, I replaced all verbs and nouns which do not occur in WordNet (e.g. proper nouns) by the string "notinwordnet". This decreases the size of the resulting LSC model significantly, since instead of estimating parameters separately for each of these words, only parameters for "notinwordnet" are estimated. This reduction saves memory and processing time. In the LSC approach, the number of classes as well as the number of learning iterations has to be fixed in advance. Following (Rooth et al. 1998), I used 35 classes and 400 iterations.

Figure 5.1 shows one of the clusters of the obtained LSC model (class c_{18}). In section 3.3, I have explained that a latent semantic class essentially comprises[10] semantically similar verbs and semantically similar nouns so that these verbs and nouns tend to co-occur with each other. The class in figure 5.1 illustrates this behaviour. This class particularly comprises verbs denoting some kind of changing ("open", "close", "change"), putting ("put", "sit", "fill"), or moving ("cross", "go"), as well

[9] Of course, as the beginning of the quote on page 120 indicates, Resnik does not deny the usefulness of lexical disambiguation as a preprocessing step of acquiring selectional preferences.

[10] Recall that the LSC approach is a soft clustering method, i.e. the membership of a verb v or a noun n in a class c is modelled by the conditional probability $p(v|c)$ or $p(n|c)$, respectively. Thus, stating that a class comprises certain verbs and nouns means that these verbs and nouns have comparably high membership probabilities for that class.

$p(c_{18}) = 0.0146$			
v	$p(v\|c_{18})$	n	$p(n\|c_{18})$
open#subj:obj/obj	0.0369	eye	0.0976
open#subj/subj	0.0294	door	0.0776
notinwordnet	0.0253	notinwordnet	0.0385
close#subj:obj/obj	0.0194	face	0.0328
put#subj:obj:pp.on/pp.on	0.0123	mind	0.0277
cross#subj:obj/obj	0.0111	mouth	0.0193
put#subj:obj:pp.in/pp.in	0.0104	window	0.0173
go#subj:pp.to/pp.to	0.0098	bed	0.0156
sit#subj:pp.on/pp.on	0.0094	table	0.0155
hit#subj:obj/obj	0.0088	chair	0.0143
fill#subj:obj/obj	0.0088	gray	0.0140
sit#subj:pp.in/pp.in	0.0073	wall	0.0133
change#subj:obj/obj	0.0070	side	0.0129
close#subj/subj	0.0067	line	0.0114
clear#subj:obj/obj	0.0063	floor	0.0113
...	...	...	...

Figure 5.1: Latent semantic class 18

as nouns denoting body parts (“eye”, “face”, “mouth”), parts of buildings (“door”, “window”, “wall”, “floor”), or pieces of furniture (“bed”, “table”, “chair”). Furthermore, entities denoted by such nouns are typically affected by events denoted by such verbs. Note that not all verbs or nouns, respectively, in a class are similar to each other. However, similar verbs and nouns tend to be grouped in the same classes.

In an LSC model, the ambiguity of a verb or a noun is reflected by its (graded) membership in several classes so that different senses of a certain word are prevalent in different classes. To illustrate this, figure 5.2 and 5.3 show two (artificially constructed) classes. Both of them include the noun “bank”. However, class c_{ex1} in figure 5.2 provides evidence for the sense denoting a geological formation (concept <bank#side> in WordNet), whereas c_{ex2} in figure 5.3 indicates the sense denoting a building (concept <bank#bank_building> in WordNet). In the same way, these classes indicate certain senses of the verbs included in them. For example, the verbs in figure 5.2 provide evidence for those senses of “climb” that indicate a movement rather than those senses that indicate some kind of increase, e.g. of prices (in WordNet, there are four senses of “climb”; two denote a movement and two an increase). This evidence implies at least a partial disambiguation of “climb”. In other words, the LSC model implicitly captures the information that the ‘movement’ senses of “climb” co-occur with the sense of “bank” represented by <bank#side>. It is just this kind of information which is needed for mapping pairs of verb–noun forms to appropriate pairs of verb–noun concepts. The only remaining step which is required to employ that information is a method to make it explicit, i.e. a method to automatically determine *which* WordNet senses a cluster favours for each of the verbs and nouns in it. The second step of my disambiguation approach addresses this task.

In this second step, I employed a method for disambiguating semantically similar words within a cluster. Resnik (1995*a*) proposes such a method. The basic idea is to compare the senses of the words

$p(c_{ex1}) = 0.008$			
v	$p(v\|c_{ex1})$	n	$p(n\|c_{ex1})$
walk#subj:pp.along/pp.along	0.01	hill	0.06
climb#subj:obj/obj	0.009	bank	0.06
reach#subj:obj/obj	0.009	shore	0.04
climb_down#subj:obj/obj	0.007	elevation	0.01
...	...	...	...

Figure 5.2: A latent semantic class containing "bank" in the 'geological formation' sense

$p(c_{ex2}) = 0.01$			
v	$p(v\|c_{ex2})$	n	$p(n\|c_{ex2})$
enter#subj:obj/obj	0.03	hall	0.04
build#subj:obj/obj	0.02	bank	0.03
leave#subj:obj/obj	0.02	school	0.02
plan#subj:obj/obj	0.006	palace	0.01
...	...	...	...

Figure 5.3: A latent semantic class containing "bank" in the 'building' sense

in a cluster and for each word select those sense(s) which is (are) closest to the senses of the other words. For example, consider the nouns in the class displayed in figure 5.1. Their similarity helps to mutually disambiguate each other. E.g. "chair" could denote a piece of furniture or a person. There are a number of nouns in the cluster (such as "bed", "table", etc.) which have senses that also refer to some furniture, i.e. senses which are close to the 'furniture' sense of "chair". Thus, these nouns provide strong evidence for that sense. In contrast, nouns which have a sense close to the 'person' sense of "chair" occur only rarely in the cluster.[11] Hence, only weak evidence is provided for that sense in the cluster. Similarly, the 'furniture' sense of "bed" is strongly supported by other nouns in the class with similar senses, whereas the 'geological formation' sense is not.

To formalise the notion of *closeness* or *similarity of senses*, Resnik (cf. (Resnik 1995*b*)) employs the hyponym/hyperonym hierarchy of WordNet. This approach is based on the assumption that semantic similarity of two concepts (or their dissimilarity, respectively) is reflected by their closeness (distance) in the hierarchy. To illustrate the idea, let us look at the hierarchical structure (which corresponds to WordNet 1.5) displayed in figure 5.4, in particular the concept <bank#side>. <mountainside> is very close to that concept, <hill> is a bit more distant, and <artefact> is located quite far away. This is in accordance with the intuition about the similarity of these concepts: <mountainside> is very similar to <bank#side>, <hill> is quite similar, and <artefact> is dissimilar. Now the question arises how a semantic taxonomy can be used to obtain a quantitative measure of semantic similarity or semantic distance. One obvious way to calculate the distance between two concepts would be to count the number of edges on the path between these concepts. However, the individual hyperonymy relations in WordNet represent different degrees of abstraction, and hence different semantic distances between the concepts they relate. For instance, the relation between

[11] For example, "face" has a sense in WordNet which denotes a person, as in "when he returned to work he met many new faces".

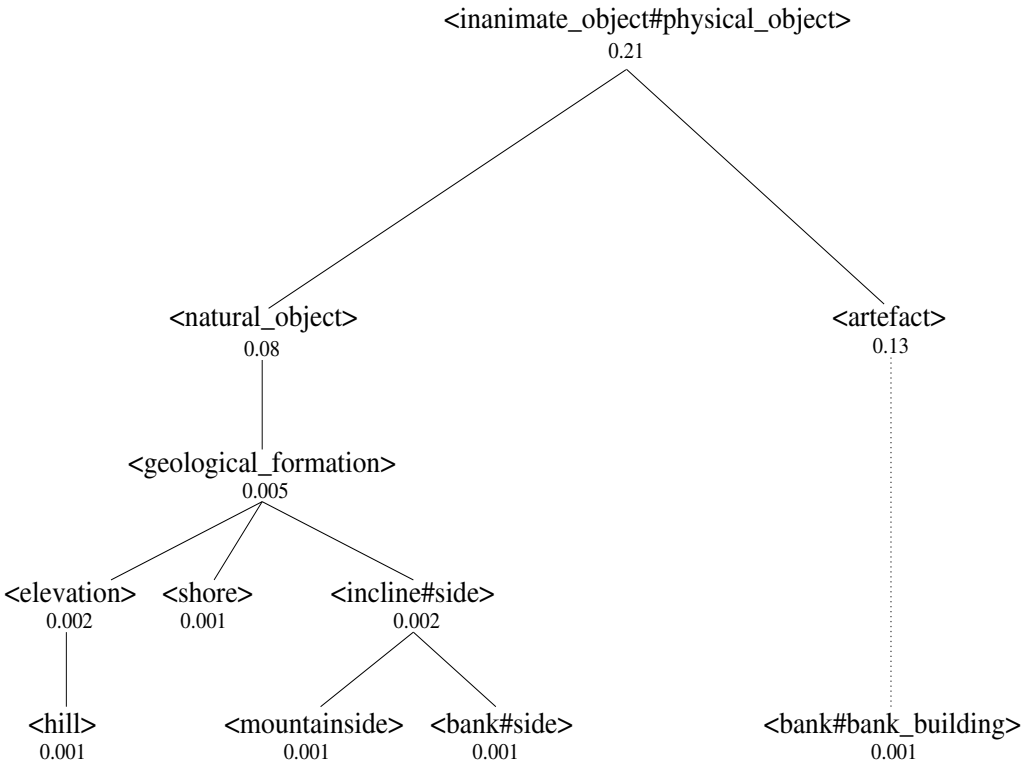


Figure 5.4: Part of the WordNet hierarchy with (artificial) concept probabilities

<natural_object> and <physical_object> obviously implies a larger abstraction step than the relation between <hill> and <elevation>. Or, to take another example, the fact that there is an intermediate node between <hill> and <geological_formation>, but there is no such node between <shore> and <geological_formation> implies that the edge-counting approach yields a distance of 2 in the former case and a distance of 1 in the latter. However, it is counterintuitive that <hill> should be twice as semantically distant to <geological_formation> as <shore>. For these reasons, Resnik's (1995*b*) definition of semantic similarity combines the structural information provided by the hierarchy with a corpus-based measure which is motivated by information theory. This measure is called *information content*. In section 3.4.1.2, I explained that in information theory, the information carried by a message is quantified by the negative logarithm of its probability. (This reflects the idea that a message which is improbable and thus highly surprising provides a larger amount of information than a message which is probable and thus has been expected.) Accordingly, the information content of a concept is defined as

$$info(cpt) = -\log p(cpt) \tag{5.8}$$

The higher the probability of a concept, the less informative (and the more abstract) it is. Figure 5.4 shows probabilities of the displayed concepts. Note that these probabilities have been artificially chosen in order to facilitate the intuitive understanding of the basic ideas explained in this section. Of course, the probabilities to be employed by the WSD approach are estimated from the training data.[12] Resnik defines the similarity of two concepts in WordNet as the amount of information which both concepts share. This is the information content of their *most informative subsumer*, i.e. the most specific superconcept that subsumes both concepts. For example, the most informative subsumer of <bank#side> and <mountainside> is <incline#side>, which has the information content $-\log 0.002 = 8.97$. The most informative subsumer of <bank#side> and <shore> is <geological_formation> with the information content $-\log 0.005 = 7.64$. (Note that the same holds for <bank#side> and <hill>. Hence, this approach overcomes the problem of the presence or absence of intermediate nodes mentioned above.) Finally, the most informative subsumer of <bank#side> and <artefact> is <physical_object> with the information content $-\log 0.21 = 2.25$. These values model the similarity of the respective concepts.

However, this similarity measure does not take into account the distance of the two compared concepts from their most informative subsumer. For example, in Resnik's approach the similarity between <bank#side> and <elevation> is equal to the similarity between <bank#side> and <geological_formation>. In both cases, the most informative subsumer is <geological_formation>, and the information content of this concept quantifies the similarity. However, it is intuitive that the latter pair of concepts exhibits a higher similarity than the former one. For this reason, I decided to adopt a different though related distance measure which takes such differences into account. This measure was defined in (Jiang & Conrath 1997).[13] In contrast to Resnik's method, it quantifies semantic distance rather than semantic similarity. This distance is defined as the difference of information content between two concepts and their most informative subsumer:

[12] This can be done by using either the word-to-concept or the word-to-sense approach. Resnik uses the former strategy, whereas I employ the latter one in my implementation.

[13] Several semantic similarity and distance measures for WordNet have been proposed. Budanitsky & Hirst (2001) review five such measures and evaluate them by comparing their performance within the semantically-driven detection and correction of real-word spelling errors. They report that the measure of Jiang & Conrath (1997) performs best.

$$
\begin{aligned}
dist(cpt_1, cpt_2) &= info(cpt_1) + info(cpt_2) \\
&\quad -2 \times info(mis(cpt_1, cpt_2)) \\
&= -\log p(cpt_1) - \log p(cpt_2) \\
&\quad +2 \times \log p(mis(cpt_1, cpt_2))
\end{aligned} \tag{5.9}
$$

where $mis(cpt_1, cpt_2)$ denotes the most informative subsumer of cpt_1 and cpt_2. The loss of information when abstracting from cpt_1 to $mis(cpt_1, cpt_2)$ is the difference of their respective information contents, i.e. $info(cpt_1) - info(mis(cpt_1, cpt_2))$. Correspondingly, the loss of information when abstracting from cpt_2 to $mis(cpt_1, cpt_2)$ is $info(cpt_2) - info(mis(cpt_1, cpt_2))$ Equation (5.9) calculates the sum of these two differences. According to this measure, the semantic distance between <bank#side> and <elevation> is $-\log 0.001 - \log 0.002 + 2 \log 0.005 = 3.66$, whereas the distance between <bank#side> and <geological_formation> is $-\log 0.001 - \log 0.005 + 2 \log 0.005 = 2.33$, i.e. the latter distance is smaller than the former, which is the desired effect.

To be applicable in Resnik's overall disambiguation approach, this distance measure has to be transformed into a similarity measure. To achieve this, I took 2 to the power of the negative distance:

$$
sim(cpt_1, cpt_2) = 2^{-dist(cpt_1, cpt_2)} \tag{5.10}
$$

For example, this measure quantifies the similarity between <bank#side> and <elevation> as 0.079 and the similarity between <bank#side> and <geological_formation> as 0.2.

In the following, I describe the algorithm which employs the similarity measure defined above to (partially) disambiguate the words in a cluster. As noted, this algorithm essentially follows (Resnik 1995*a*). However, it provides an extension to deal with soft clusters, i.e. with graded cluster memberships of words. Figure 5.5 shows a pseudo-code. The algorithm is applied separately to each class c and, within a class, separately for verbs and nouns.[14] For each word w_i, it estimates the probability distribution $p(sns_k(w_i)|w_i, c)$ of senses $sns_k(w_i)$ given w_i and c. To achieve this, a pairwise comparison of all the words (verbs or nouns, respectively) is performed. For each word pair (w_i, w_j), the similarity between each sense of w_i and each sense of w_j is computed. Those senses $sns_{k1}(w_i)$ and $sns_{k2}(w_j)$ which are closest to each other, i.e. which yield the maximal similarity value between the two words provide a certain amount of evidence for each other. Essentially, this evidence is quantified by the similarity of the two senses itself. Thus, the more similar they are, the more evidence they provide for each other. However, the graded membership of the words in the class also have to be taken into account. The idea is that the higher the membership probability of w_j in c, the more evidence its sense $k2$ provides for the sense $k1$ of w_i, and conversely, the higher the membership probability of w_i in c, the more evidence its sense $k1$ provides for the sense $k2$ of w_j. Therefore, the evidence the similarity $sim(sns_{k1}(w_i), sns_{k2}(w_j))$ provides for $sns_{k1}(w_i)$ is weighted (multiplied) by $p(w_j|c)$, and the evidence this similarity provides for $sns_{k2}(w_j)$ is weighted by $p(w_i|c)$.[15] For sense $k1$ of w_i, a support counter is incremented by the amount of evidence computed in this way. The same is done for sense $k2$ of w_j. The support counter for a specific sense of a word quantifies

[14] Actually, the algorithm examines all verbs or all nouns, respectively, in the training data, since each verb and each noun has a certain membership probability for each class (which may be very close to 0).

[15] (Resnik 1995*a*) does not deal with soft clusters with a graded membership of words. Therefore, he does not employ a weighting of this kind.

Input:
– a set of words $W = w_1, \ldots, w_n$ comprising all verbs *or* all nouns in the training data
– the corresponding set of membership probabilities $p(w_1|c), \ldots, p(w_n|c)$ for a certain class c

for i = 1 to n, j = 1 to $i-1$ {
  for $k1$ = 1 to $num\_senses(w_i)$, $k2$ = 1 to $num\_senses(w_j)$ {
    if $sim(sns_{k1}(w_i), sns_{k2}(w_j)) = \max sim(sns_x(w_i), sns_y(w_j))$ {

      $evidence = sim(sns_{k1}(w_i), sns_{k2}(w_j)) \times p(w_j|c)$
      increment $support_{i,k1}$ by $evidence$
      increment $normalisation_i$ by $evidence$

      $evidence = sim(sns_{k1}(w_i), sns_{k2}(w_j)) \times p(w_i|c)$
      increment $support_{j,k2}$ by $evidence$
      increment $normalisation_j$ by $evidence$
    }
  }

for i = 1 to n {
  for k = 1 to $num\_senses(w_i)$ {
    if $normalisation_i > 0$ {
      $p(sns_k(w_i)|w_i, c) = \frac{support_{i,k}}{normalisation_i}$
    } else {
      $p(sns_k(w_i)|w_i, c) = \frac{1}{num\_senses(w_i)}$
    }
  }
}

Output: probabilities $p(sns_k(w_i)|w_i, c)$ for $w_i \in W$, $k = 1, \ldots, num\_senses(w_i)$

Figure 5.5: Pseudo-code of the algorithm for disambiguating words in an LSC cluster

how much this sense is supported by the other words in the cluster. In addition, for each word a normalisation counter adds up all evidence for *any* sense of that word. Finally, after every possible word pair has been checked, the support count for each sense $sns_k(w_i)$ is divided by the normalisation counter, which contains the sum of support counts for all senses of w_i. This normalisation transforms the sense support counts into sense probabilities $p(sns_k(w_i)|w_i, c)$. These probabilities are returned. For those words for which no evidence for certain senses can be found, the method falls back to the uniformity hypothesis, i.e. a uniform distribution of the possible senses is assumed.

When applied to the nouns in class c_{ex1} in figure 5.2, the algorithm examines, for instance, the two words "bank" and "elevation". Among all possible combinations of senses of these words, the concepts <bank#side>, which is sense 1 of "bank", and <elevation> (hyponym of <geological_formation>), which is sense 4 of "elevation", are closest. As mentioned above, their similarity is 0.079. The class membership probability of "bank" is 0.06, that of "elevation" is 0.01. Therefore, the support counter for sense 1 of "bank" is incremented by $0.079 \times 0.01 = 0.00079$, and the support counter for sense 4 of "elevation" is incremented by $0.079 \times 0.06 = 0.00474$. Other nouns will provide evidence for other senses of these words. However, such nouns tend to have comparably low probabilities in the class so that this support will be low as well. For example, among all senses of "bank", <bank#bank_building> (sense 9) is closest to the concept representing the noun "storehouse" (this noun is monosemous in WordNet). Imagine that "storehouse" has a membership probability of 10^{-7} in c_{ex1} and the similarity of <bank#bank_building> and <storehouse> is 0.09. Then the evidence that "storehouse" provides for sense 9 of "bank" amounts to $0.09 \times 10^{-7} = 9 \times 10^{-9}$.

The algorithm in figure 5.5 yields estimates of the probabilities $p(sns_i(v)|v, c)$ and $p(sns_j(n)|n, c)$, i.e. the conditional probabilities of a certain sense of a word (a verb or a noun, respectively) given that word and a certain class. With these probabilities and the parameters of the LSC model (i.e. $p(c)$, $p(v|c)$, and $p(n|c)$), we can estimate the joint probabilities $p(sns_i(v), sns_j(n))$ of noun and verb senses:

$$p(sns_i(v), sns_j(n)) = \sum_c p(c) \times p(sns_i(v)|c) \times p(sns_j(n)|c) \quad (5.11)$$

$$= \sum_c p(c) \times p(sns_i(v)|v, c)p(v|c) \times p(sns_j(n)|n, c)p(n|c) \quad (5.12)$$

With this joint probability, we can compute the conditional probability of the co-occurrence of a verb sense and a noun sense given the underlying verb and noun form:

$$p(sns_i(v), sns_j(n)|v, n) = \frac{p(sns_i(v), sns_j(n))}{p(v, n)} \quad (5.13)$$

which is the ratio of the probabilities defined in equation (5.11) and equation (3.5) on page 59, respectively. This probability estimation in turn is used to estimate verb–noun sense pair frequencies according to equation (5.4) on page 119.

It should be noted that word sense disambiguation on the one hand, and learning selectional preferences on the other hand are interdependent tasks. Disambiguating the verbs and nouns in the training data as a preprocessing step (as proposed in this section) aims at improving the acquisition of selectional preferences. Conversely, we have noticed (cf. sections 1.1.1 and 2.2.1) that information of

selectional patterns are useful for word sense disambiguation. In section 3.4, we have seen that several researchers who developed a method for learning selectional preferences have evaluated the performance of their approach in a WSD task. Therefore, an obvious strategy to disambiguate the training data would be to make use of information provided by the acquired selectional preferences themselves. This strategy was pursued by McCarthy (1997). She adopted the Li and Abe approach. In an initial step, she learned selectional preferences assuming uniform sense distributions of nouns. Then, in a second step, she employed these preferences for disambiguation by assigning a noun the sense with the highest preference value. This is a reasonable approach. However, I decided to make use of a disambiguation strategy which does not rely on that technique of selectional preference acquisition which it is intended to improve, in order to avoid the amplification of weaknesses (i.e. tendencies of certain errors and biases) of this technique.

5.3 Transforming the WordNet Structure

As noted several times, WordNet deviates from the structure of the noun hierarchy assumed for the tree cut approach in two respects. Firstly, word senses are not only represented by leaves, but by all nodes. Secondly, WordNet does not have a pure tree structure, but is a DAG, i.e. a concept can have multiple parents. This section deals with the transformation of the WordNet structure to overcome these inadequacies.

To handle the first issue, two strategies have been proposed in the literature. Li and Abe themselves solve the problem by pruning the hierarchy: If a node represents a word sense that occurs in the data, then all subtrees of that node are removed so that this node becomes a leaf. Thus, the hierarchy is pruned at those nodes that represent the most general word senses occurring in the sample. If a word sense corresponds to a concept which is located in a discarded subtree, then this word sense is represented by the root of this subtree. For instance, if the sample $S_{assassinate}$ contained a number of words which are hyponyms of <person> and a single instance of "being", then the hierarchy is pruned at the node <life_form#organism#being#living_thing>, which now captures the instance "being" and the hyponyms of <person> in the data. Due to this pruning, the learning algorithm only can select among cuts which are at least as general as <life_form>, which, as we have seen, is above the appropriate generalisation level. This example illustrates that this strategy is inadequate for our needs, since some rare usages of the examined verb (or noisy data) might have the effect that the appropriate generalisation level is "pruned away".

As I already mentioned in section 4.1.1.1, I adopted another strategy, which to my knowledge was first proposed by McCarthy (1997) and Abney & Light (1998). To shortly recapitulate this strategy: The idea is to (virtually) create for each inner node an additional node that represents a sense of the words which belong to the synset of that node. This additional node becomes a hyponym of the original node. In this way, all word senses are captured by leaf nodes.

The frequency counts for these leaves are calculated straightforwardly. The starting point is the estimation of sense probabilities given pairs of verb–noun forms, which was derived in section 5.2. Recall equation (5.4) on page 119:

$$freq(sense_i(v), sense_j(n)) = freq(v, n) \times p(sense_i(v), sense_j(n)|v, n)$$

This equation maps frequencies of verb–noun form pairs to frequencies of verb–noun sense pairs via

the estimation of conditional sense probabilities given word forms.

Each word sense corresponds to exactly one synset. Within our construction of the noun hierarchy, the leaf concepts capture the senses. Thus, each noun synset at a leaf position represents those senses that correspond to that synset. For verbs, the correspondence between a sense and its synset is even more straightforward: As we do not generalise over verb concepts (cf. the discussion in section 3.4.4), there is no need to alter the verb hierarchy (nor to treat the verbs as part of a hierarchy at all). Thus, there are no duplicated verb concepts, and each verb synset represents a sense of each verb which is a member of it. To get the frequency count of a synset that represents word senses, the counts of the corresponding senses are just added up. We have to distinguish two cases here. To calculate the frequency of a noun (leaf) concept $nsns$ co-occurring with a particular verb concept $vcpt$, we have to take into account all verb senses corresponding to $vcpt$ as well as all noun senses corresponding to $nsns$:

$$freq(vcpt, nsns) = \sum_{\substack{sense_i \equiv vcpt \\ sense_j \equiv nsns}} freq(sense_i, sense_j) \tag{5.14}$$

For example, <grow#develop> represents sense 8 of "grow" and sense 13 of "develop". <pupil#schoolchild> represents sense 3 of "pupil" and sense 1 of "schoolchild". Thus, $freq(<\text{grow\#develop}>,<\text{pupil\#schoolchild}>) = freq(sense_8(grow), sense_3(pupil)) + freq(sense_8(grow), sense_1(schoolchild)) + freq(sense_{13}(develop), sense_3(pupil)) + freq(sense_{13}(develop), sense_1(schoolchild))$.

The counts acquired according to equation (5.14) are employed to learn a tree cut model that represents the selectional preferences of $vcpt$, i.e. by the procedure Find-Assoc-MDL (cf. section 3.4.3.4). In contrast, the procedures Find-MDL and Calc-p-Closure, which calculate a probability distribution over noun concepts regardless of a particular verb, require concept counts over the total sample. For a leaf noun concept $nsns$, the overall frequency is obtained by summing its co-occurrence frequencies for all verb concepts $vcpt$ in the sample:

$$freq(nsns) = \sum_{vcpt} freq(vcpt, nsns) \tag{5.15}$$

The second problem mentioned at the beginning of this section (which McCarthy (2001) handily calls the "DAG issue") requires a virtual transformation of the WordNet DAG structure to a pure tree structure. *Virtual* transformation means that the structure is not really altered, but the hierarchy is processed in a way that simulates a tree structure. Two parts of the learning approach are involved in this simulation:

1. the propagation of noun leaf concept counts to higher concepts (a preprocessing step)
2. the traversal of the structure by the learning algorithm

In section 3.4.3.4 we saw that the learning algorithm recursively traverses the structure top-down. A side-effect of this processing is that the DAG structure is "resolved" into a tree structure. Nodes that have multiple parents are processed multiple times, once for each parent. For example, as <person>

is a hyponym of both <life_form> and <causal_agent>, this concept (and thus its hyponyms) is processed twice, once as a child of <life_form>, and once as a child of <causal_agent>. In this way, a "virtual copy" of such a node (and its descendants) is created for each of its parents, and the DAG is "broken into a tree" as illustrated in figure 5.6 (a virtual copy is indicated by a dashed link). Hence, one crucial part of transformation is done already by the acquisition algorithm and does not require any modification.

The other crucial part is the calculation of the concept frequency counts. As mentioned in section 3.4.3.2, the tree cut approach employs the word-to-sense method to obtain concept frequencies, i.e. the frequencies of word senses are propagated to all their ancestors in the hierarchy, and for each concept, the frequencies accumulated at it add up to its count. In fact, there are several possibilities of how to perform this propagation. Following Ribas' approach explained in section 3.4.2.1, the frequency of a concept is the sum of the frequencies of the word senses which are subsumed by that concept. Taking into account the restriction that all senses are leaves, equation (3.22) on page 71 can be reformulated as

$$freq(cpt) = \sum_{sns:\, is\_leaf(sns) \wedge sns \nearrow^* cpt\}} freq(sns) \tag{5.16}$$

where $is\_leaf(sns)$ denotes the condition that sns is a leaf. (It should be pointed out that this equation counts each leaf node only once, not taking into account "virtual copying" of leaves due to the DAG-to-tree transformation.) If the hierarchy is a proper tree structure, then this is equivalent to the sum of the frequencies of the immediate hyponyms (i.e. the children) of the concept:

$$freq(cpt) = \sum_{cpt_{child} \in children(cpt)} freq(cpt_{child}) \tag{5.17}$$

However, if the hierarchy is a DAG, then the equations (5.16) and (5.17) yield different values, as we will see below.

Employing either of the two equations for calculating concept counts has its advantages and drawbacks. Both possibilities have been proposed. Li and Abe obtain concept frequencies according to equation (5.17). The advantage of this alternative is that the counts are consistent with the tree structure which is virtually created by the top-down processing of the hierarchy. The duplication of subtrees is reflected by the corresponding counts. In this way, the transformation of WordNet into a tree structure is completed. Figure 5.6 illustrates this approach. As can be seen, the frequency count for a concept is the sum of the counts of the children (e.g. for <entity>, the count is $200 + 105 = 305$).

The drawback of this approach is that multiplying certain subtrees amounts to multiplying the portion of those items in the sample which are covered by that subtree. For example, as the concept <person> is processed twice, all instances in the sample denoting a person are counted twice. In particular, these multiple counts accumulate in the counts of those concepts which subsume nodes with common descendants. Concerning the probability distribution to be acquired, this results in a bias towards such concepts, since larger probabilities are estimated for them. For example, the frequency of the top node <entity> contains the count of <person> twice, which results in a higher probability estimated for <entity>.

(McCarthy 2001) calculates concept frequencies according to equation (5.16). Thus, the frequency of

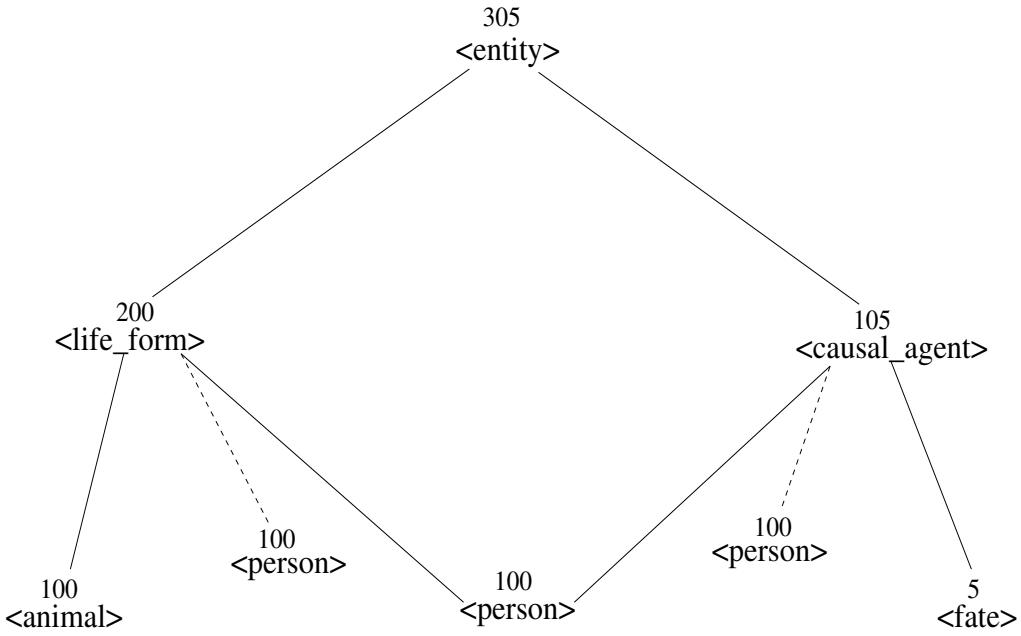


Figure 5.6: Breaking a DAG into a tree structure according to equation (5.17)

a concept is the sum of the frequencies of all senses subsumed by it, as described in section 3.4.2.1. Taking into account the constraint that the senses are located at the leaves, this means that the count for each leaf is immediately propagated to all its ancestors. To demonstrate that this approach may yield different results than the previous one, let us apply it to the hierarchy depicted in figure 5.6. For the sake of simplicity, let us assume that this is the complete hierarchy we deal with, i.e. that <animal>, <person>, and <fate> are the leaves. Then the counts for <life_form> and <causal_agent> are the same as in figure 5.6. <life_form> subsumes the leaves <animal> and <person>, thus its count is $100 + 100 = 200$. Likewise, <causal_agent>, subsuming <person> and <fate>, gets the count $100 + 5 = 105$. However, the count for <entity> is different: this concept subsumes all three leaves so that its count is $100 + 100 + 5 = 205$ (as opposed to 305 according to equation (5.17)). This corresponds to the real amount of data captured by the hierarchy in this example.

The advantage of this approach is that the count of a a concept captures each sense it subsumes once and only once. Thus, the count of an *individual* concept is not biased by the duplication of sense counts. However, the *total* count of the sample as captured by a tree cut model may differ for different cuts. If a cut contains two or more concepts which have a common subconcept, then the count for this subconcept forms part of the count of each of its ancestors on the cut. In the example discussed in the previous paragraph, a cut along the concepts <life_form> and <causal_agent> comprises a total count of $200 + 105 = 305$, whereas a cut at the top node <entity> represents a count of 205 (which is appropriate w.r.t. the data). McCarthy explicitly refers to that inconsistency. She nonetheless justifies her approach by claiming that it minimises the negative impact of applying the tree cut learning algorithm to the WordNet DAG structure. She notes that multiple hyperonymy is very rare in WordNet; it occurs at less than 1% of the concepts, mostly at low levels of the hierarchy.

For my work, I decided to develop a different approach for retrieving concept frequency counts, in order to avoid the drawbacks which the two previously sketched alternatives bring about. This decision is motivated by the desideratum pointed out in section 1.3 that the methods devised in this thesis should be language-independent, i.e. applicable to any language for which the required resources (a lexical semantic net and a parsed corpus) are available. While multiple hyperonymy is rare in WordNet, this is not necessarily the case in wordnets for other languages. For example, in GermaNet (cf. (Hamp & Feldweg 1997), (Kunze & Wagner 2001)), cross classification of concepts via multiple hyperonymy is a major structuring principle and thus very common. Actually, 11.5% of the GermaNet concepts have more than one parent. This illustrates that an approach which handles the DAG issue in a more principled way is crucial w.r.t. language-independence.

The general idea of the method for concept frequency estimation which I propose here is as follows: As in the work of Li and Abe, the count of a concept is directly determined by the counts of its children (i.e., its immediate hyponyms). This simulates a tree structure. However, a concept does not necessarily inherit the *total* count from each of its children. If a concept has multiple parents, then the count of that concept is divided among its parents. In this way, counts are not duplicated, and thus no bias towards certain parts of the sample is created. The frequency portion that a concept cpt_{child} passes to each of its parents is determined by a probability distribution $p(cpt|cpt_{child})$ where cpt is a parent of cpt_{child}. Thus, the frequency of a concept is given by

$$freq(cpt) = \sum_{cpt_{child} \in children(cpt)} freq(cpt_{child}) \times p(cpt|cpt_{child}) \tag{5.18}$$

The crucial question is how to estimate the distribution $p(cpt|cpt_{child})$. I decided to guide this estimation by correlating the frequencies of a concept's parents: The count of a concept is apportioned

among its parents according to their respective frequency, relative to the frequencies of the other parents. To be more exact, for a concept cpt, the distribution $p(cpt_{parent}|cpt)$ is estimated by the ratio of the frequency of cpt_{parent} and the sum of the frequencies of all parents of cpt:

$$p(cpt_{parent}|cpt) = \frac{freq(cpt_{parent})}{\sum_{cpt' \in parents(cpt)} freq(cpt')} \tag{5.19}$$

In the trivial case in which cpt has only one parent, $p(cpt_{parent}|cpt)$ is 1, i.e. the complete concept frequency is propagated to that parent. I will provide an informal motivation of equation (5.19) below.

It is easy to notice that the equations (5.18) and (5.19) depend on each other. The probability of the parent given a child concept in equation (5.18) is estimated by equation (5.19), whereas the parent frequencies in equation (5.19) are obtained by equation (5.18). Therefore, to make these equations applicable, it is necessary to assume certain initial values for either the concept frequencies or the parent probabilities and then to re-calculate these quantities by applying (5.18) and (5.19) several times (iterations). There is no obvious way to stipulate plausible initial concept counts. However, it is quite straightforward to initialise the parent probabilities by assuming uniform distributions:

$$p(cpt_{parent}|cpt) = \frac{1}{|parents(cpt)|} \tag{5.20}$$

In this way, the count of a concept is equally apportioned to its parents in the initial iteration. As the parents of a concept have different additional children, this iteration yields different counts for them. Thus, in the following iterations, equation (5.19) will estimate differing probabilities for the parents of a concept. Note also that in general, an iteration step changes the counts and probabilities. The approach proposed here can be viewed as an EM-style algorithm where equation (5.18) corresponds to the E-step and equation (5.19) to the M-step. For my experiments, I performed only one further iteration after the initial one.

An example of this approach is shown in figure 5.7. Here, the initialisation step equally apportions the count of <person> to its two parents; each parent inherits the count $\frac{100}{2} = 50$. Then, in the re-estimation step, the <person> count is divided relative to the frequencies of the parents: <life_form> inherits $100 \times \frac{150}{150+55} = 73.17$, while <causal_agent> receives $100 \times \frac{55}{150+55} = 26.83$ from <person>. (The counts for <animal> and <fate> are completely propagated to their respective parents.) Note that the count for the top node <entity> does not change. It corresponds to its unbiased frequency in the data.

This strategy of splitting the concept count before propagation to multiple parents requires a corresponding adjustment of frequencies when the learning algorithm processes the hierarchy top-down (cf. point 2. on page 131). For example, in step 9. of the procedure Find-Assoc-MDL (figure 3.16 on page 85), the description lengths of two tree cut models are compared: the cut located at the root of the processed subtree and the optimal cut below the root. The counts of the concepts on the lower cut (which are needed to calculate the data description length) have to be accommodated so that only that portion is taken into account that has been propagated to the concept on the upper cut, i.e. the root of the processed subtree. Thus, for each concept cpt_i on the lower cut, the quantity $freq(cpt_i)$ (written as $\sharp(t_i, S_v)$ in the pseudo-code) has to be replaced by $p(cpt_r|cpt_i) \times freq(cpt_i)$ where cpt_r is the root of the subtree. Let $cpt_{k_1}, \ldots, cpt_{k_n}$ be the hyperonymy chain of concepts between cpt_i and cpt_r

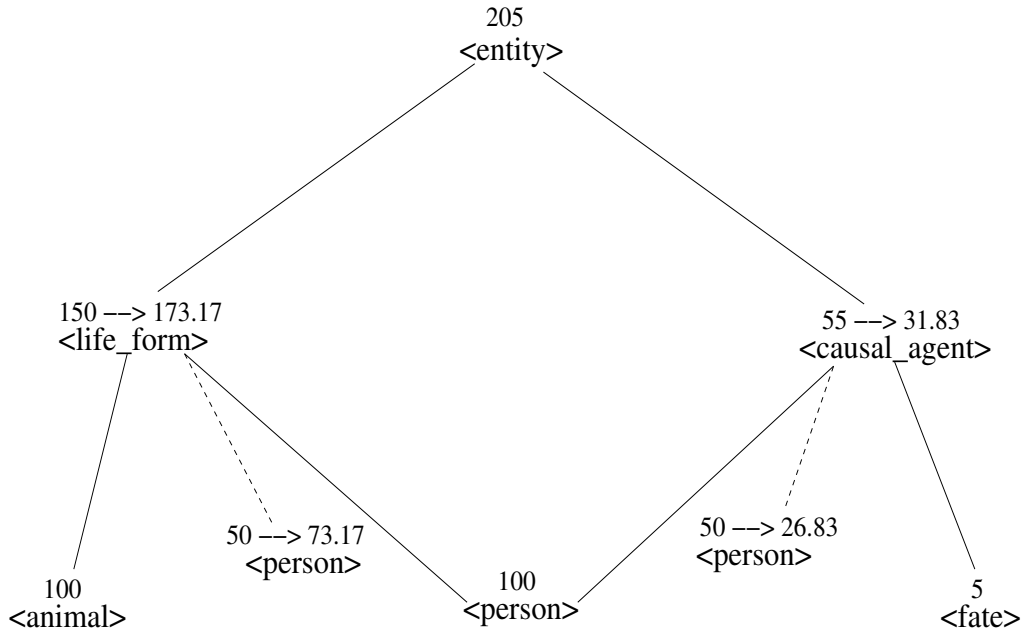


Figure 5.7: Re-estimating frequencies

in the processed subtree.[16] Then $p(cpt_r|cpt_i)$ is given by the following equation:[17]

$$p(cpt_r|cpt_i) = p(cpt_{k_1}|cpt_i) \times p(cpt_{k_2}|cpt_{k_1}) \times \ldots \times p(cpt_{k_n}|cpt_{k_{n-1}}) \times p(cpt_r|cpt_{k_n}) \quad (5.21)$$

Similar changes are necessary for the procedures Find-MDL (step 8.) and Calc-p-closure (step 6.).

A possible intuitive access to the general idea that the count of a concept is divided among its parents might be to understand hyperonymy in a more subjective manner: Instead of "is a kind of", a hyperonymy relation could be interpreted as "is perceived / referred to as". This means that multiple hyperonyms represent different aspects of a concept which might have different salience. For example, a person might be primarily referred to as a life form in some situations, and as a causal agent in other situations. The probabilities $p(<\text{life\_form}>|<\text{person}>)$ and $p(<\text{causal\_agent}>|<\text{person}>)$, together with the corresponding split of the count of <person>, model the relative salience of these two aspects w.r.t. <person>. The way I propose to estimate these probabilities employs the only empirical quantitative information about the parent concepts that is accessible: their total frequency. A parent that has a high frequency (compared to the other parents) gets a high probability, while a parent with a (comparably) low frequency is assigned a low probability. The count of a parent concept reflects its “global” salience; the comparison with the counts of the other parents reflects peculiarities of their common child.

More formally, the approach described here can be viewed as performing a hierarchical soft classification of noun senses. The concepts can be viewed as soft classes of senses, and multiple hyperonymy corresponds to graded membership. For example, all instances of <person> are graded members of both <life_form> and <causal_agent>. The degree of membership is represented by $p(<\text{life\_form}>|<\text{person}>)$ and $p(<\text{causal\_agent}>|<\text{person}>)$, respectively.

Furthermore, I would like to point out that there are certain analogies between this approach and the idea proposed by Abney and Light to treat the WordNet structure as a Hidden Markov Model (cf. section 3.4.6). If, as in their approach, the nodes of the hierarchy are interpreted as states and the links between them as arcs of an HMM, then the propagation of counts corresponds to transitions in this HMM so that the concept counts $freq(cpt)$ calculated by equation (5.18) can be viewed as the expected number of transitions from state cpt and the probabilities $p(cpt_{parent}|cpt)$ calculated by equation (5.19) can be interpreted as transition probabilities (usually written as a_{ij} in the HMM framework) from state cpt to cpt_{parent}. However, there are crucial differences between the strategy sketched here and the approach of Abney and Light. Firstly, Abney and Light aim at modelling selectional preference as an integrated stochastic process. In contrast, the strategy described in this section has a much more modest purpose, namely assigning concept frequencies in a sound and plausible way in order to obtain a virtual tree structure of the WordNet hierarchy. Consequently, the demands for the Abney and Light approach are much more sophisticated. For example, Abney and Light expected from their model to select the correct senses of the word forms in the corpus, a requirement that gave rise to many difficulties, as noted in section 3.4.6. For my approach of propagating concept counts, this is not an issue, because word sense disambiguation is handled by a different preprocessing module (described in the previous section). Secondly, the hierarchy is processed in opposite

[16] Since the top-down processing virtually transforms the hierarchy into a tree structure, this chain is always unique for the actual point of traversal.

[17] It is easy to keep track of this product during recursive processing. Thus, this modification does not affect the complexity of the algorithm.

directions in the two approaches. In the HMM of Abney and Light, the root node is the initial state and the transitions are directed downwards, i.e. from a parent to a child, until a leaf (a final state) is reached. However, approaches to collect concept frequencies from word frequencies naturally start at the leaves[18] and propagate counts upwards. Thirdly, although the parent probabilities $p(cpt_{parent}|cpt)$ can be viewed as analogous to the transition probabilities of an HMM, they are acquired in different ways: A transition probability a_{ij} is estimated by the fraction

$$a_{ij} = \frac{\text{expected number of transitions from state } i \text{ to } j}{\text{expected number of transitions from state } i} \tag{5.22}$$

In contrast, equation (5.19), reformulated in terms of the HMM framework, would be

$$a_{cpt\ cpt_{parent}} = \frac{\text{expected number of transitions from state } cpt_{parent}}{\sum_{cpt' \in parents(cpt)} \text{expected number of transitions from state } cpt'} \tag{5.23}$$

The exact relationship between equation (5.22) and equation (5.23) has to be further investigated. Here, I leave it at the concluding statement that the approach proposed here resembles the HMM approach of Abney and Light at first glance, but looking more closely at the two strategies reveals substantial differences.

In (Wagner 2003), I report experiments which compare the impact of the frequency propagation approach proposed here and the approach proposed by Li and Abe (i.e. employing equation (5.17) on page 132). These experiments acquire tree cut models to represent selectional preferences in the way I proposed in chapter 4. I separately employed each of the two concept frequency estimation approaches and compared the acquired tree cut models. I performed the learning algorithm on German data and GermaNet, because, as mentioned above, the structure of the GermaNet hierarchy deviates from a pure tree structure to a large extent. These experiments show that the results retrieved employing the two approaches differ significantly and the bias introduced by the ad-hoc strategy proposed by Li and Abe is considerable.

5.4 A Gold Standard for Evaluation

To evaluate (i.e. assess the quality of) an NLP technique, there are basically two alternative ways. One alternative is to measure the performance of the technique when employing it for a particular application. This is the way which has commonly been used to evaluate approaches for learning selectional preferences so far. Specifically, the approaches described in section 3.4 have been evaluated by applying them for word sense disambiguation (cf. (Resnik 1997), (Ribas 1995*b*), (Agirre & Martinez 2002), (Abney & Light 1998)), PP attachment disambiguation (cf. (Resnik 1993), (Abe & Li 1996)), or other tasks (e.g. the prediction of object omissibility in (Resnik 1993) or a pseudo-disambiguation task[19] in (Clark & Weir 2002)). Evaluating one or more approaches w.r.t. a specific task reveals no more and no less than the strengths and weaknesses of these approaches regarding that task. This may be viewed

[18] Within the HMM interpretation, the leaves would be the initial states which emit word senses, while the inner nodes are non-emitting states.

[19] This pseudo-disambiguation task (cf. footnote 20 on page 92) illustrates that the exercise employed for evaluation is not necessarily a task required in a "real life" NLP application, but might instead be based on an artificial setting.

as an advantage or a disadvantage, depending on the actual aim of learning selectional preferences. If they are acquired in order to be utilised in a particular application, then an evaluation with regard to this application is most appropriate.[20] However, the primary goal of this thesis is to develop an approach for learning selectional preferences to enrich lexical-semantic resources, in particular wordnets. Since wordnets are designed as general-purpose lexical resources, this preference information is intended to be application-independent, i.e. to be employed in various applications. Therefore, an application-independent evaluation is highly desirable here.

The second basic alternative for evaluating an NLP technique is to compare the results induced by that technique with a gold standard, i.e. a set of "correct" data of the type that the technique induces or predicts. This evaluation method has been applied for a variety of tasks, e.g. POS tagging, parsing, or word sense disambiguation. Employing a gold standard is a method of *directly* evaluating an NLP approach, in contrast to the *indirect* method of testing its performance in some application. Furthermore, evaluation by a gold standard is in a sense more generic, since it is not biased towards a specific application. Therefore, as just pointed out, this alternative is more adequate for our purposes. The main problem with this evaluation approach is to obtain an appropriate gold standard. In general, any supervised learning method is usually evaluated using a gold standard (the test data), which comprises a comparably small subset of the available supervised data (whereas the majority of these data is used for training). For unsupervised learning methods, data which are suitable to serve for this purpose are not available per se. In fact, as the data feeding a gold standard have to be created (or at least corrected) manually, compiling these data is time- and labour-intensive. Thus, the need for reusable gold standards for different tasks, which are available to researchers in the field, has become obvious. Several efforts have been made to compile reusable test data (as well as training data for supervised approaches) for specific NLP tasks. For example, the PennTreebank (cf. (Marcus et al. 1993)) has commonly been employed for the evaluation of parsers (cf. also (Grishman, Macleod & Sterling 1992)). For other tasks, the creation of training and test data has been embedded in initiatives to organise a competitive evaluation of NLP techniques, e.g. SENSEVAL for word sense disambiguation (cf. (Killgarriff 1998)) or MUC for information extraction (cf. (Grishman & Sundheim 1996)).

For the task of learning selectional preferences to be included in WordNet, an appropriate gold standard should comprise a set of thematic role relations between verb and noun concepts of WordNet. Since WordNet does not provide relations of this type, such a gold standard is not immediately available. However, as stated in section 1.2.3 and 2.3.1, the formal specification of EuroWordNet allows for a certain inventory of thematic role relations, and indeed some of the monolingual wordnets in EWN (the wordnets for Dutch, English, Estonian, Italian, and Spanish) do contain role relations. These relations have been manually encoded or extracted from other lexical resources, respectively. I employed them for compiling a gold standard by mapping them to WordNet. Thanks to the structure of EWN, this mapping can be done straightforwardly. The inter-lingual index (ILI) plays a key role here: As mentioned in section 1.2.2, the ILI essentially consists of the concept nodes of WordNet 1.5. Furthermore, (almost) each concept in a monolingual wordnet in the EWN database is related to one or more ILI concepts, in most cases via a synonymy or a near-synonymy relation, encoding semantic (near-) equivalence between the monolingual concept and the ILI concept. Thus, the desired gold standard can be retrieved in a simple way: Extract from the wordnets in EWN all thematic role relations where both the source node and the target node are linked to an ILI concept via a (near-)synonymy relation. For each of these relations, replace the source and the target concept by the corresponding ILI con-

[20] Obviously, this argument does not apply to artificial tasks like pseudo-disambiguation.

role type	#
AGENT	2253
PATIENT	1017
INSTRUMENT	2107
LOCATION	220
DIRECTION	29
SOURCE_DIRECTION	40
TARGET_DIRECTION	68
RESULT	423
unspecified	594
all	6751

Table 5.1: Number of different types of role relations mapped to WordNet

cept(s).[21] In section 1.2.3, I listed some role relations from the Italian wordnet, together with the ILI (i.e. WordNet) concepts corresponding to the involved Italian concepts. I repeat these relations here to illustrate this translation procedure:

(5.24) <abbeverarsi> INVOLVED_AGENT <animale>
<drink> *<animal>*

(5.25) <camminare> INVOLVED_AGENT <pedone>
<walk> *<pedestrian>*

(5.26) <lessare> INVOLVED_PATIENT <cibo>
<boil> *<food>*

(5.27) <sbaraccare> INVOLVED_PATIENT <cosa>
<remove> *<object#inanimate_object>*

(5.28) <lavare> INVOLVED_INSTRUMENT <acqua>
<wash> *<water>*

The set of relations created by this method comprises 6751 relations altogether. Table 5.1 lists the number of relations for each role type. Table 5.2 shows some further examples of these mapped relations. Note that I did not mention the RESULT relation type in section 2.3.1.[22] The reason for this is that this type is not covered by the EuroWordNet deliverable (Alonge 1996) which serves as the official documentation of the different relation types in EWN and their semantics. The RESULT relation will not play a role in this work.

Looking more closely at the relations obtained by the procedure sketched above, it turns out that not all of them are suited for a gold standard to evaluate the task of this thesis. Some relations,

[21] In those cases where a monolingual concept is linked to multiple ILI concepts, the mapping procedure creates a corresponding number of different gold standard relations.

[22] Henceforth, I will often skip the prefix INVOLVED_ in the role relation label for the sake of brevity.

verb concept	relation type	noun concept
<crouch#stoop#bow>	AGENT	<person>
<adore>	AGENT	<admirer#adorer>
<adore>	AGENT	<believer#worshipper>
<teach>	AGENT	<teacher#instructor>
<crow>	AGENT	<cock#rooster>
<change>	PATIENT	<clothing#clothes#wear>
<scalp#remove_the_scalp>	PATIENT	<person>
<edit#cut>	PATIENT	<writing#written_material>
<pick_up#receive>	INSTRUMENT	<antenna>
<pick_up#receive>	INSTRUMENT	<radio#tuner>
<pick_up#receive>	INSTRUMENT	<television>
<see>	INSTRUMENT	<spectacles#glasses>
<see>	INSTRUMENT	<eye#oculus>
<cure#heal>	LOCATION	<sanatorium>
<drive#motor>	DIRECTION	<road_route>
<disembark#debark#go_ashore>	SOURCE_DIRECTION	<vessel#craft>
<hop_on#mount#climb_on#bestride>	TARGET_DIRECTION	<horse>
<teach>	RESULT	<cognition#knowledge>
<worry>	unspecified	<concern#worry#headache>

Table 5.2: Examples of role relations mapped to WordNet

e.g. <delouse> INVOLVED_PATIENT <louse> or <silt> INVOLVED_PATIENT <sediment>, involve incorporated arguments. Since these arguments are not realised as syntactic complements, such relations cannot be induced by analysing verb–complement pairs extracted from corpora. Encoding incorporated arguments by role relations in EWN is explicitly intended. Alonge (1996, p. 32) states: "...our links will encode (only) data on the semantic features of arguments *incorporated* in the meaning of a verb, which certainly determine also the kind of arguments which a verb allows as its complements, but which do not exactly coincide with them." Thus, role relations in EWN might represent incorporated arguments or semantic information about overt arguments. Note that this corresponds to Jackendoff's approach to model incorporated arguments in the same way as selectional restrictions on syntactic complements (cf. section 2.2.2).

Other relations like <address> INVOLVED_PATIENT <addressee> indicate a noun concept that itself is perfectly adequate, but does not (in terms of the hierarchy) subsume the majority of the noun instances which really co-occur with the respective verb in the corpus. Any noun referring to a human could occur as the patient of <address>. Thus, the expected and desired behaviour of the learning algorithm would be to generalise to the <human> level. In contrast, <addressee> is a subconcept of <human> which has no hyponyms itself. Such relations involve a noun concept that characterises an argument of the respective verb *intensionally*. However, the learning approach proposed here aims at acquiring a concept that subsumes all preferred concepts, i.e. provides an *extensional* characterisation of the argument in question. In principle, it makes sense to encode thematic relations employing intensional concepts. However, such relations cannot be derived by generalising from corpus instances. They could rather be acquired by examining derivational patterns. To allow for a fair evaluation of acquisition approaches examined in this thesis, relations of the kinds mentioned in this and the previous

paragraph have to be removed from the gold standard.

Another weakness of the gold standard is that in some cases it is slightly contradictory regarding the generalisation level of the nouns. For example, there are two INVOLVED_PATIENT relations for the verb <own#have#possess>: one to <asset> (originating from the Dutch wordnet) and one to <possession> (originating from the English wordnet). The former concept is a hyponym of the latter one, i.e. the two concepts deviate by one level of abstraction in the hierarchy. Such cases indicate that humans have a rough idea of the appropriate abstraction level for representing selectional preferences of a given verb, but human intuition about which generalisation level is *exactly* adequate in this respect is not quite clear-cut. This insight fits in a general experience which is prevalent in computational (and traditional) lexicography, namely that in the field of lexical semantics things are not as clear-cut as for other tasks such as POS tagging or parsing. In my view, the vagueness of human intuition regarding appropriate generalisation justifies an analogous interpretation of the results acquired by the learning algorithm: In the evaluation experiments described below and in chapter 7, I will count those acquired relations as "hit" (i.e. as successfully learned) which deviate from the corresponding gold standard relations by 0 or 1 hierarchy levels, i.e. where the acquired noun concept either exactly matches the gold standard noun concept or is an immediate hyponym/hyperonym of it. A fair evaluation should not postulate that the accuracy and consistency of an automatic method exceeds the accuracy and consistency of human intuition.

Apart from these particular problems, a general shortcoming of the gold standard constructed in the way described above is the involved translation step. Generally, the translation from the individual (non-English) wordnets to WordNet may introduce some inaccuracies due to deviating selectional preferences of corresponding verbs in different languages, differing hierarchical structures of the language-specific wordnets, or—of course—plain errors. It should be pointed out that a certain amount of errors is likely to occur in any manually built resource, and hence in any gold standard. However, this amount usually is small so that the respective resources are sufficiently accurate to serve for evaluation tasks. Concerning the gold standard extracted from EWN, I have manually inspected a considerable portion of the obtained relations (especially the PATIENT relations, see below), and I regard the actual amount of translation inaccuracies as tolerably low.

5.5 Experiments

This section describes the experiments I carried out to evaluate different approaches for learning selectional preferences. These experiments compare the acquired preferences to the gold standard just described. In chapter 3, I argued that the tree cut approach proposed in (Abe & Li 1996) is best suited for the task of this thesis. Therefore, the evaluation mainly focuses on comparing the performance of the method of Abe & Li (1996) (*standard MDL*) and the modification of this approach by introducing a weighting factor (*weighting*), which I have proposed in section 4.2. Both variants compute the preference values (association norms) in the same way. Furthermore, both variants determine the appropriate level of generalisation by the MDL principle (in the broader sense). They just differ in how they employ the MDL principle. To round up the picture, I also examined a completely different approach, namely the approach of Ribas (cf. section 3.4.2). Ribas uses a different formula to compute the preference value and a completely different strategy to determine the adequate generalisation level (a simple greedy selection algorithm). Anyhow, in contrast to other strategies introduced in chapter 3, Ribas' approach meets the formal desiderata which are relevant for learning thematic role relations: firstly, the formula of the preference value defines a natural distinction between preference

and dispreference; secondly, the set of selected noun concepts is nonredundant, i.e. it does not contain concepts which are hyponyms/hyperonyms of each other, and thus this set defines a certain level of abstraction. Hence, this approach in principle is suitable for learning thematic role relations.

5.5.1 Setting

For the experiments, I employed the training data introduced in section 5.1. As noted several times, employing these syntactic data to learn thematic roles requires linking syntactic complements to thematic roles. I will propose a strategy to fulfil this task in the next chapter. In chapter 7, I will report extensive evaluation experiments which include that strategy. The experiments reported here are intended as a case study, in order to assess the suitability of the above-mentioned approaches for learning thematic role relations. Therefore, I will not systematically address the linking problem here, but instead concentrate on one specific thematic role, namely Patient. The choice of this particular role is connected with the simple assumption that Patients usually are syntactically realised as objects. Following this assumption, I extracted all verb–object pairs from the data and used them as the basis for extracting PATIENT relations.

A manual inspection of the PATIENT relations in the gold standard showed that the simplifying heuristic to link the grammatical object to the Patient role turned out to be useful in general. Most of the PATIENT relations in the gold standard identify preferences for the objects of the respective verbs. However, a certain amount of PATIENT relations were inappropriate here. As mentioned in section 5.4, some of these relations encode incorporated arguments, while others provide an intensional characterisation of the Patient argument. Such information is not inducible by generalising from corpus instances. A third problematic group, which is specific to the Patient role, comprises unaccusative verbs (e.g. <ascend>), which realise their Patient (e.g. <cloud>) as subject. For the experiments I describe here, I excluded these kinds of relations from the gold standard. The remaining set contains 662 relations for 368 verb concepts altogether.

To test the impact of the disambiguation approach proposed in section 5.2, I applied the learning algorithm once on non-disambiguated data (dividing the evidence for a word uniformly among its senses, cf. section 5.2.1) and once on disambiguated data (applying the method described in section 5.2.2). In both cases, I propagated the frequency counts of the noun senses to their superconcepts following the method proposed in section 5.3.

In the experiments described below as well as in the evaluation experiments reported in chapter 7, I employed the formula for optimising the parameter description length introduced in section 4.1.1.1 (equation (4.1) on page 99). However, in contrast to the experiments described in chapter 4, I did not employ a threshold that prunes the search space by discarding concepts with very low probabilities. Initial experiments within the setting outlined here revealed that employing a threshold significantly decreases performance.

Another peculiarity of the experiments presented here and in chapter 7 is that I calculate the frequencies and probabilities of noun concepts regardless of a particular verb (i.e. $freq(ncpt)$ and $p(ncpt)$) by taking into account all nouns in the training data, rather than only those nouns that occur in the examined syntactic or semantic relation (here, the object or, respectively, the PATIENT relation). This deviates from the experiments in chapter 4 as well as from the most approaches described in section 3.4. The alternative I pursue here is explicitly addressed by Ribas (1995*b*) (cf. footnote 7 on page 67). He points out that this variant also takes into account the information (in other words, the semantic

number (percentage) of gold standard noun concepts	standard MDL		weighting; $C =$ 1000		10 000	
exactly matched	12	(3.6%)	108	(32.2%)	147	(43.9%)
matched by 1 level hyponym	2	(0.6%)	34	(10.1%)	55	(16.4%)
matched by 1 level hyperonym	18	(5.4%)	49	(14.6%)	50	(14.9%)
matched by $>= 2$ level hyponym	1	(0.3%)	8	(2.4%)	14	(4.2%)
matched by $>= 2$ level hyperonym	152	(45.4%)	90	(26.9%)	36	(10.7%)
not matched	150	(44.8%)	46	(13.7%)	33	(9.9%)

Table 5.3: Comparison of tree cut models acquired from non-disambiguated data with the gold standard

preferences) that a certain type of argument slot provides about the nouns which fill that slot, irrespective of a concrete verb. In section 6.5, I will discuss (and utilise) the finding that thematic role types provide very general semantic characterisations of their fillers (e.g. Agents tend to be animate beings, while Instruments are usually inanimate objects. These role-specific, but verb-independent characterisations are reflected by a preference acquisition approach if and only if the preference value indicates the difference between the noun concept distribution given a certain verb and a certain role type on the one hand and the overall noun concept distribution independent from verbs *and* from role types on the other hand.

5.5.2 Results

5.5.2.1 Tree Cut Approach: Non-Disambiguated Data

For both variants standard MDL and weighting (with different values of the constant C), I compared the noun concepts acquired for a verb concept with the corresponding noun concepts in the gold standard. To avoid problems introduced by sparse data, I selected those verb concepts which have a frequency count of at least 50 for evaluation. For the non-disambiguated data, 174 verbal concepts met this criterion. Table 5.3 shows the number and the percentage of the noun concepts in the gold standard which were exactly matched, not matched at all, or matched by more general or more specific concepts in the tree cut model. Dispreferred concepts, i.e. concepts with a preference value < 1, are considered as not matching the gold standard.

While the results for the standard MDL algorithm are not satisfying, the results for the weighting algorithm are very promising. For $C = 10\ 000$, 43.9% of the noun concepts are exactly matched by the tree cut model (as opposed to 3.6% with standard MDL). If one also takes the approximate matches (1 level deviation) into account, then the matching rate is 75.2% (vs. 9.6% for standard MDL). In section 5.4, I have argued that it is appropriate to count approximate hits, because human intuition about the *exact* generalisation level is not always clear-cut either. This is illustrated by the finding that the gold standard itself contains cases where a verb concept is related to two noun concepts which deviate by one level of abstraction, i.e. where one is a hyponym of the other.

As can be seen, the standard MDL approach tends to learn too general concepts or not to match

verb	weighting model	standard MDL model
<pump#raise_with_a_pump>	<gas>	<entity>
<start#start_up#set_in_motion>	<engine>	—
<climb#climb_up#mount#go_up>	<road#route>	<artifact#artefact>
<send#direct>	<mail>	<relation>
<tame#chasten#subdue>	<animal#beast#creature#fauna>	<entity>
<buy#purchase#take>	<commodity#goods>	<commodity#goods>
<cook#change_by_heating>	<food#nutrient>	<food#nutrient>
<suppress#inhibit#subdue#curb>	<idea#thought>	<psychological_feature>
<record#tape>	<material>	<abstraction>
<operate#control>	<device>	—
<pick#pluck#cull>	<flower#bloom#blossom>	<entity>
<plug_in#connect >	<device>	—
<cultivate#foster_the_growth_of>	<plant#flora#plant_life>	—

Table 5.4: Some gold standard concepts exactly matched by weighting tree cut models ($C = 1000$) and the corresponding concepts in the respective standard MDL tree cut models

preferred concepts at all, respectively. This tendency is illustrated by table 5.4. This table shows some examples of gold standard concepts which are exactly matched by the respective tree cut models acquired by the weighting algorithm ($C = 1000$), and the corresponding concepts in the respective tree cut models acquired by the standard MDL algorithm. (A dash indicates that a gold standard concept is not matched by the model.)

The best performance is achieved with a rather high value of C, and hence, with a rather low generalisation level of the acquired preferences. The reason for this is the heterogeneous nature of the gold standard. The gold standard is rather inconsistent w.r.t. the degree of generalisation: On the one hand, very general concepts, e.g. <own> INVOLVED_PATIENT <possession>, on the other hand, very specific concepts, e.g. <add> INVOLVED_PATIENT <appendix>, have been encoded. It is obvious that specific concepts are captured by a cut at a low generalisation level. More surprisingly, it turned out that in many cases, low-generalisation cuts also capture rather general concepts, e.g. <person>. This is due to the treatment of inner nodes sketched in section 5.3. It is often the case that a specific cut contains virtual leaves, which represent senses of those words which characterise the corresponding inner nodes. For example, the virtual leaf that corresponds to the node <person> represents a sense of the word "person". If a cut contains this leaf and if the corresponding preference score exceeds 1, then "person" co-occurs with the verb under consideration with a significant frequency in the training sample so that the algorithm recognises the corresponding concept as preferred. In other words: in such cases, general concepts are acquired due to immediate evidence from the corpus. For these reasons, a bias to low-generalisation cuts yields the best overall performance.

Note that this evaluation focuses on recall, i.e. the percentage of relations in the gold standard which are captured by the learned model. Unfortunately, the gold standard is far from being "exhaustive"; in general, it does not represent *all* noun concepts which are preferred by a given verb concept. Furthermore, information about which noun concepts are *dispreferred* by a certain verb concept is not available. Thus, it is not possible to carry out a quantitative evaluation of precision, i.e. the percentage

number (percentage) of gold standard noun concepts	standard MDL		weighting; $C =$ 1000		10 000	
exactly matched	19	(8.0%)	66	(27.8%)	78	(32.9%)
matched by 1 level hyponym	5	(2.1%)	27	(11.4%)	45	(19.0%)
matched by 1 level hyperonym	25	(10.5%)	31	(13.1%)	28	(11.8%)
matched by >= 2 level hyponym	1	(0.4%)	8	(3.4%)	13	(5.5%)
matched by >= 2 level hyperonym	115	(48.5%)	41	(17.3%)	23	(9.7%)
not matched	72	(30.4%)	64	(27.0%)	50	(21.1%)

Table 5.5: Comparison of tree cut models acquired from disambiguated data with the gold standard

of learned thematic relations which are "really correct".

Manual inspection shows that the acquired selectional preferences contain a considerable amount of noise. For example, the gold standard contains a relation between the verb <hound#hunt#trace> and the noun <game> (an animal hunted for food or sport). The tree cut model for <hound#hunt#trace> correctly models preference for this concept, but, additionally, preference for three other senses of "game", e.g. a hyponym of <competition#contest>. Of course, this kind of noise is due to the fact that the data are non-disambiguated.

5.5.2.2 Tree Cut Approach: Disambiguated Data

For the evaluation using disambiguated data, I again selected the verb concepts with a frequency of at least 50. This selection yielded a set of 122 verbal concepts. Table 5.5 shows the evaluation results.

Compared to learning from non-disambiguated data, the percentage of exactly or approximately (at level 0 or 1) matched concepts is worse for the weighting algorithm, e.g. 63.7% for $C = 10\ 000$ (as opposed to 75.2%). The reason for this is that disambiguation errors misinform the learning algorithm to a certain degree, as they favour "incorrect" senses, which results in dropping "correct" noun concepts, which occur in the gold standard. Analogously, some of the "correct" verb senses were dropped. For this reason, fewer verb concepts in the gold standard are captured by the data. However, the results are still promising. Improved approaches to WSD should decrease this effect.

As expected, disambiguation reduces noise in the tree cut models. For example, two of the three erroneous senses of "game" (s.a.) are not modelled as preferred concepts of <hound#hunt#trace> any more. Unfortunately, there is no way to measure automatically how much noise and how much "good" data is dropped by the disambiguation step. For $C = 10\ 000$, disambiguation yields 560 preferred noun concepts per verb concept on average, as opposed to 1326 concepts without disambiguation. For $C = 1000$, disambiguation reduces this number from 378 to 242. Thus, the relative overall reduction of preferred concepts (58% or 36%, respectively) is much higher than the loss of accuracy w.r.t. the gold standard. This indicates that noise reduction outweighs the loss of useful information caused by the disambiguation step. To get a better idea about this loss of "correct" information, a detailed manual comparison of the cuts with and without disambiguation would be necessary. Anyway, in a semi-automatic setting in which the algorithm learns candidate concepts which are manually inspected afterwards, reducing the "candidate space" without losing too many valid candidates is very useful.

number (percentage) of gold standard noun concepts	non-disambiguated data		disambiguated data	
exactly matched	39	(11.6%)	35	(14.8%)
matched by 1 level hyponym	43	(12.8%)	31	(13.1%)
matched by 1 level hyperonym	53	(15.8%)	36	(15.2%)
matched by $>= 2$ level hyponym	17	(5.1%)	20	(8.4%)
matched by $>= 2$ level hyperonym	149	(44.5%)	69	(29.1%)
not matched	34	(10.1%)	46	(19.4%)

Table 5.6: Comparison of the preferences acquired by Ribas' approach with the gold standard

Standard MDL performs better with disambiguation, but the results are still unsatisfying (20.6% exact or approximate matches). Here, for most of the verbs, the tree cut model contains rather general concepts, for which the danger to be "disambiguated away" is low.

5.5.2.3 Ribas' Approach

To round up the picture, I also evaluated the performance of the preference acquisition approach of Ribas (cf. section 3.4.2) w.r.t. the gold standard. I ran Ribas' learning algorithm on the non-disambiguated and the disambiguated data mentioned above. To obtain the preference scores (cf. equation (3.25) on page 72), I used the same probability estimates as for computing the preference values in the experiments described above.[23] Table 5.6 shows the obtained results. Analogously to the tree cut experiments, concepts with a negative association score are considered as not matching the gold standard.

There is not much difference between the results using non-disambiguated or disambiguated data. The (exact or approximate) matching rate is 40.2% and 43.1%, respectively. These rates are clearly below the performance of the weighting approach, for both examined values of C. But, after all, they are significantly higher than the matching rates for standard MDL. The latter point is remarkable regarding the fact that Ribas' approach uses a simple heuristic to find the adequate level of generalisation, whereas Li and Abe emphasise that their approach uses a theoretically well-motivated method to determine the appropriate abstraction level.

For a semi-automatic setting, which includes manual inspection of the acquired concepts, the number of learned preferences is important. For Ribas' approach, the average number of acquired preferred concepts per verb is 638 for non-disambiguated data and 318 for disambiguated data. These numbers lie between the corresponding values of the weighting approach for $C = 10\ 000$ (1326 and 560, respectively) and $C = 1000$ (378 and 242, respectively). Thus, for $C = 1000$, the weighting algorithm

[23] Note that there is a difference between Ribas' approach and the tree cut approach regarding the general probabilities of noun concepts regardless of a particular verb ($p(ncpt)$). Ribas calculates all probabilities by simple Maximum Likelihood Estimation. However, Li and Abe estimate $p(ncpt)$ by a tree cut model. I decided to employ the latter alternative here, because this allows a comparison between the different preference acquisition approaches which is not biased by using different probabilities. Note also that there is no corresponding difference between the verb-specific probabilities $p(ncpt|vcpt)$ for the two approaches, because for the concepts on the learned tree cut model, these probabilities are computed by MLE as well.

acquires a smaller amount of preferred concepts, but still achieves a higher recall rate.

5.5.3 Preliminary Conclusion and Further Proceeding

The experiments sketched here show that the standard MDL approach proposed by Li and Abe is hardly suitable for learning thematic role relations. However, the weighting approach proposed in chapter 4 is promising. Up to 75% of the acquired concepts were appropriate (recall). This rate was retrieved by using non-disambiguated data. The disambiguation strategy introduced in section 5.2 significantly reduces noise, but also reduces the rate of correctly acquired concepts (although this reduction is rather moderate). This happens because some correct senses are "disambiguated away". The results for non-disambiguated data indicate the "potential" of the approach. Improved disambiguation techniques should eliminate noise caused by incorrect senses, but keep the correct ones.

In section 5.3, I introduced a new approach for transforming the WordNet DAG structure into a tree structure. In (Wagner 2002), I report experiments which I carried out using the same data and strategies as for the experiments reported in section 5.5.2.1 and 5.5.2.2, but employing the ad-hoc strategy for the structural transformation proposed by Li and Abe. It turns out that in both studies, the results are very similar. This finding is in accordance with the fact that multiple hyperonymy, i.e. the deviation from a pure tree structure, is very rare in WordNet. Thus, the transformation strategy I proposed does not matter much here. This is different for wordnets where multiple hyperonymy is much more common, e.g. GermaNet (cf. (Wagner 2003)).

As noted, the current gold standard is rather heterogeneous w.r.t. the degree of generalisation. Therefore, it seems reasonable to divide the relations in the gold standard into subgroups according to the generalisation level of the involved noun concepts and employ such subgroups separately for evaluation. I will examine this possibility in section 7.8.

To evaluate my approach for learning thematic role relations on a broader basis, it is necessary to examine other thematic relations as well. To achieve this, a comprehensive linking strategy is required. I will address this issue in the next chapter.

Chapter 6

Mapping Syntactic Arguments to Thematic Roles

The basic learning scenario adopted for this thesis is the acquisition of thematic role relations from parsed corpora. As noted several times, this setting involves a fundamental gap which has to be bridged in one or another way: Whereas the training data contain information about the *syntactic* arguments of verbs, the learning task is to acquire preferences for the *semantic* arguments of verbs. The problem which has to be solved is that a certain semantic role type can be realised by several syntactic argument types, and conversely, a certain syntactic argument type can realise several semantic role types. In section 5.1, I illustrated this fact by referring to the examples (5.1)–(5.3). Here, I repeat these sentences again as (6.1)–(6.3), and add (6.4) and (6.5):

(6.1) The jealous husband broke the window.

(6.2) A hammer broke the window.

(6.3) The jealous husband broke the window with a hammer.

(6.4) The furious wife broke the TV with a chair.

(6.5) The TV broke.

These sentences show that the Instrument can be realised as subject (6.2) or *with*-PP ((6.3); (6.4)), and the Patient can be expressed as object ((6.1)–(6.4)) or subject (6.5). On the other hand, the subject can realise an Agent ((6.1); (6.3); (6.4)), an Instrument (6.2), or a Patient (6.5).

In section 2.1, I described the problem of linking syntactic arguments to thematic roles from a theoretical linguistic perspective. Within the linguistic research summarised in that section, the primary question could be stated as follows: "Given a particular sentence, which thematic role underlies a certain syntactic argument?" For the learning task of this thesis, the problem has to be viewed from a slightly different perspective. Here, the primary question has to be reformulated: "Given a particular verb, which syntactic arguments realise a certain thematic role of that verb (in the data)?" In other words, the linking task in our context is not to assign appropriate roles to the arguments in concrete sentences. Rather, the task is to group together arguments (in differing syntactic configurations) which are fillers of a specific thematic role of the verb under consideration. For example, regarding the verb

"break", the nouns "husband" and "wife" correspond to the Agent, "window" and "TV" to the Patient, and "hammer" and "chair" to the Instrument. When such role-specific noun groups are established, the learning approach for selectional preferences can be separately applied to each of them. The preferred concepts on the tree cut acquired from a role-specific noun collection are strong candidates for being connected to the respective verb concept via a corresponding role relation. For example, if the Instrument group yields a cut containing <artefact> as a preferred concept, this suggests an INVOLVED_INSTRUMENT relation between <break> and <artefact>.

In this chapter, I will present a strategy to acquire such role-specific argument groups for the verb under examination. This strategy employs the training data I described in section 5.1. As described in that section, the data I use consist of items which represent syntactic and lexical information about sentences in the corpus. Formally, a data item is a pair where the left component contains

- a verb (the main verb of a sentence)
- a subcategorisation frame (the list of syntactic complement types of the verb in that sentence)
- a particular argument slot captured by the subcategorisation frame (a particular complement type occurring in the sentence)

while the right component contains the noun which is the head of the particular complement type encoded in the left component. Notationally, a data item has the form *verb#frame/slot noun*. Thus, a sentence like (6.3) would be represented in the training data by the following items:

```
break#subj:obj:pp.with/subj  husband
break#subj:obj:pp.with/obj  window
break#subj:obj:pp.with/pp.with  hammer
```

For example, the first item above contains the information that there is a sentence in the training corpus in which "break" occurs with the subcategorisation frame *subj:obj:pp.with*, i.e. with a subject, an object, and a PP with preposition "with", and that the subject is "husband". As a data item represents a syntactic argument of a verb, the task of obtaining role-specific argument noun groups effectively involves building groups of data items.

6.1 Overall Strategy

The strategy I propose for mapping syntactic arguments to thematic roles consists of three stages:

1. build groups of syntactic argument types such that arguments which realise the same thematic role are assembled in the same group
2. apply heuristics to determine which thematic role types correspond to which argument groups
3. apply additional (heuristic) semantic filters to cope with concurrent role assignments and to eliminate noise

The first stage investigates which syntactic argument types are likely to realise the same thematic roles. Here, the term "syntactic argument type" refers to a certain grammatical function (argument slot) and a complete subcategorisation frame to which this grammatical function belongs. Note that this information is encoded in the left component of the training data items. Henceforth, I will use the term *frame-argument configuration* to refer to such types. Examples of frame-argument configurations (in the notation used in the data items) are *subj:obj/obj* (a direct object), *subj:obj/subj* (a transitive subject), or *subj/subj* (an intransitive subject). The sub-task addressed in stage 1 is to group these configurations according to which semantic role(s) they might express. In general, the resulting groups will differ for different verbs. Consider the following sentences:

(6.6)

a. John ate the apple.

b. John ate.

c. *The apple ate.

(6.7)

a. John closed the door.

b. The door closed.

c. *John closed.

For both verbs "eat" and "close", the transitive subject expresses the Agent and the direct object the Patient (cf. (6.6 a) and (6.7 a)). However, for "eat", the intransitive subject expresses the Agent (as in (6.6 b)), whereas for "close", the intransitive subject realises the Patient (as in (6.7 b)). Hence, for "eat", the intransitive and the transitive subject have to be grouped together, while for "close", the intransitive subject has to be grouped with the object.

The example sentences under (6.6) and (6.7) suggest a way to cluster frame-argument configurations adequately. The basic idea is to group those frame-argument configurations together that have similar "fillers", i.e. similar nouns at the corresponding argument position. Examples (6.6) illustrate that for "eat", the nouns which occur as the intransitive subject coincide with the nouns occurring as the transitive subject, but not with the nouns instantiating the object. Conversely, examples (6.7) show that for "close", the fillers of the intransitive subject correspond to the fillers of the object, but not the ones of the transitive subject. This motivates the basic idea to cluster the frame-argument configurations exhibited by the verb under examination according to the similarity of the nouns they tend to relate to that verb. I will describe a principled statistical approach to perform this clustering automatically, employing the training data.

Each of the groups acquired in this way contains frame-argument configurations that agree in the semantic role(s) they express, but at this point, it is not determined *which* roles correspond to which argument groups. This kind of information cannot be immediately acquired from the data; that would require a training corpus where thematic roles are explicitly annotated (cf. section 6.6). However, the properties of certain frame-argument configurations in a cluster allow to formulate rules for determining corresponding thematic role types (or, at least, excluding inappropriate role types). Such linking rules can be straightforwardly obtained from the linguistic insights sketched in section 2.1. For example, all three linguistic theories I mentioned in that section state that if an Agent is expressed in a

sentence, then it must be expressed by the subject. From this fact, one can derive the rule that a group containing an object type cannot correspond to the Agent role. In this way, grouping frame-argument configurations constrains the possible role assignments to individual configurations. For instance, an intransitive subject can in general express an Agent or a Patient. But if this type is grouped together with the object type, then it cannot realise an Agent. Hence, building argument groups allows to apply linking rules to frame-argument configurations to which they could not be applied if they were taken into account in isolation. Actually, employing a statistical clustering method and linguistically motivated linking rules is a way of combining quantitative and symbolic evidence. As we will see, the linking rules I will propose generally are heuristic in nature. In the 2. stage of the approach described in this chapter, such heuristics are applied to each group to recognise the corresponding role types.

From the argument groups labelled with role types, it is straightforward to obtain role-specific groups of noun complements: Together with the examined verb, a frame-argument configuration forms the left component of several training data items. Thus, all those data items (pertaining to the examined verb) whose left component belongs to a group labelled with a specific role type form a role-specific collection of data items. Therefore, the nouns at the right component of the data items in such a collection constitute a role-specific multiset of complements of the examined verb. The algorithm for acquiring selectional preferences is separately applied to each of these noun complement groups to acquire thematic role relations of the respective types. For example, suppose we inspect the verb "eat" and yield (among others) an argument group containing the frame-argument configurations *subj/subj* and *subj:obj/subj* that is labelled with the role type AGENT. Then all nouns occurring in data items whose left component is *eat#subj/subj* or *eat#subj:obj/subj* are collected and the learning algorithm is applied to the resulting set to acquire INVOLVED_AGENT relations.

Unfortunately, stages 1 and 2 sketched above do not yield unique role assignments in all cases. For example, as illustrated in (6.1) and (6.2), a transitive subject might express an Agent or an Instrument. If, for example, a group just comprises the transitive subject type, then the syntactic linking rules applied in stage 2 are not able to distinguish between the two possible roles. In general, if a frame-argument configuration is ambiguous w.r.t. the role it might express and there is no other frame-argument configuration in the same group that resolves this ambiguity, then the group receives multiple role assignments. Furthermore, if an argument group is assigned a unique role, but contains an ambiguous frame-argument configuration, then the role-specific noun group obtained from that argument group as described above may contain noise, i.e. some noun complements which express another role. For instance, a noun group classified as AGENT could include several nouns expressing a Location (which are realised as subject).

Fortunately, one can formulate simple heuristic semantic filters which are able to separate different roles in ambiguous noun groups as well as eliminate noise. For example, to distinguish between Agent and Instrument one can employ the heuristic that Agents usually are animate beings, while Instruments are normally artefacts. Such heuristics are very general semantic characterisations, in a sense semantic preferences, of thematic role types. Essentially, these preferences are of the same kind as the preferences which are acquired by the learning algorithm discussed in the previous chapters, though the former are at a very high level of generalisation, being valid independently of a particular verb. Therefore, a natural way to employ such heuristics is to relate them to the selectional preferences acquired by the learning algorithm. A straightforward possibility to implement this idea is to apply these heuristics as filters of the acquired tree cut models. This is done in stage 3 mentioned above. In other words, stage 3 is applied *after* applying the learning algorithm, not—like stage 1 and 2—before. For illustration, suppose that stages 1 and 2 yield (among others) a collection of data items which is labelled with the two role types AGENT and INSTRUMENT. The preference acqui-

1. Cluster semantically similar syntactic arguments **(stage 1 of linking strategy)**
2. Apply heuristics to assign thematic role types the clusters **(stage 2 of linking strategy)**
3. Apply WSD approach to the data
4. Propagate frequencies of word senses to higher concepts in the concept hierarchy
5. Apply MDL-based learning algorithm to acquire tree cut models that represent selectional preferences
6. Apply semantic filters to these tree cuts **(stage 3 of linking strategy)**

Figure 6.1: Steps of the overall approach for learning thematic role relations proposed in this thesis. The stages of the linking strategy are explicitly marked.

sition algorithm is run on the nouns in these data items and acquires a tree cut model that represents the preferences for both roles Agent and Instrument. After that, stage 3 applies the above-mentioned heuristics to filter those concepts from the cut which correspond to each of these roles: Those concepts which denote an animate being (i.e. concepts which are subsumed by <person>, <animal>, or <causal_agent>) are filtered as candidates for the INVOLVED_AGENT relation, while those concepts that denote an artefact (i.e. concepts subsumed by <artefact>) are extracted as candidates for the INVOLVED_INSTRUMENT relation.

To clarify the embedding of the three stages of the linking strategy I propose in my overall approach of learning thematic role relations, figure 6.1 provides a brief overview of all steps of this approach and the order of their application. The components of the linking strategy are highlighted.

Overall, I would like to emphasise that the separation of the three sub-tasks as sketched in the previous paragraphs is motivated by two reasons. Firstly, this proceeding makes best use of the available information presented earlier in this thesis. In particular, it turns out that the LSC model which I use to lexically disambiguate the data (cf. section 5.2) also provides the information which enables grouping arguments together which realise the same role (stage 1). Furthermore, heuristics for determining role types based on the linguistic research concerning argument linking (cf. section 2.1) and knowledge about possible diathesis alternations can be straightforwardly formulated and applied to these argument groups (stage 2). Finally, semantic preferences on certain role types, which as well are motivated by linguistic research (cf. section 2.2.2) can be easily employed as filters of the learned tree cut models (stage 3).

Secondly, this strategy modularises the language- and theory-neutral aspects on the one hand and the language- or theory-specific aspects on the other hand. The clustering approach I propose for stage 1 is independent from both a particular language and a particular theory of thematic roles.[1] However, the heuristics I employ for stage 2 depend to a large extent on the peculiarities of English. For other

[1]The linguistic interpretation I offer in section 6.2.2 is plausible, but not constitutive for motivating the approach.

languages, different heuristics would have to be formulated. (This immediately becomes obvious if one considers that these heuristics include certain prepositions as indicators of certain roles, e.g. "with" for Instrument vs. "on" for Location vs. "onto" for Goal.) Furthermore, the formulation of role assignment rules depends on the adopted inventory of roles and their characterisations. Finally, the semantic filters applied in stage 3 are significantly influenced by the adopted role inventories and definitions as well. (The rule that an Agent is animate presupposes that nonvolitional Agents are excluded by the definition of the Agent role, or at least treated as exceptions.) As I emphasised in section 1.3, the approach for learning thematic role relations should be as language-independent as possible in order to be suitable for wordnets in any language. Furthermore, it preferably should be compatible to different linguistic role theories and inventories, to be applicable for the various existing proposals in that area. The isolation and delimitation of the necessary language- and theory-specific parts of the strategy for learning thematic role relations facilitate its adaption to different languages and theoretical assumptions.

This chapter is organised as follows: Section 6.2 explains the method which I use to group syntactic arguments and discusses a preliminary experiment to test this method (stage 1). In section 6.3, I propose and motivate some straightforward heuristics for determining appropriate role types for the individual argument groups (stage 2). In section 6.4, I sketch how the labelled argument groups are employed to prepare the input to the algorithm for learning selectional preferences. Section 6.5 describes the semantic filters I use to further discriminate role types after selectional preference acquisition (stage 3). Finally, section 6.6 relates the linking strategy proposed in this chapter with related approaches in the literature. Comprehensive experiments employing this strategy will be described in chapter 7.

6.2 Creating Role-Specific Groups of Syntactic Arguments

The step described in this section consists in forming groups of frame-argument configurations which are associated with a certain verb. In terms of the training data, the task is to cluster those types of left components which refer to the verb under consideration. For example, consider the syntactic arguments of the verb "break". The left components in the data items representing the sentences (6.1) and (6.2) above are

```
break#subj:obj/subj
break#subj:obj/obj
```

while the syntactic structure of the sentences (6.3) and (6.4) is represented by the following left components:

```
break#subj:obj:pp.with/subj
break#subj:obj:pp.with/obj
break#subj:obj:pp.with/pp.with
```

and the structure of sentence (6.5) is represented by

```
break#subj/subj
```

These six frame-argument configurations have to be grouped according to the thematic roles they represent. The roles involved here are Agent, Patient, and Instrument. Therefore, one group for each of these three roles is required. The Patient is realised as the object in (6.1)–(6.4) and as intransitive subject in (6.5). Thus, the corresponding group consists of the two frame-argument configurations which encode the object and the one encoding the intransitive subject:

```
break#subj:obj/obj
break#subj:obj:pp.with/obj
break#subj/subj
```

The Agent is represented as the subject in (6.1), (6.3), and (6.4), and not expressed at all in (6.2) and (6.5). Hence, the corresponding group comprises the frame-argument configurations

```
break#subj:obj/subj
break#subj:obj:pp.with/subj
```

The Instrument is realised as a subject in (6.2) and as a *with*-PP in (6.3), and (6.4), and not expressed in (6.1) and (6.5). Thus, the corresponding group contains the frame-argument configurations

```
break#subj:obj/subj
break#subj:obj:pp.with/pp.with
```

Note that the argument *break#subj:obj/subj* occurs in two groups. This reflects the fact that the subject in a transitive sentence without a *with*-PP might represent an Agent or an Instrument. This ambiguity of that frame-argument configuration introduces noise in both noun groups corresponding to the argument groups which it is a member of. Since some nouns attached to this configuration express an Agent and some an Instrument, but all of these nouns are included in the two noun groups in question, the group representing Agents contains some nouns expressing an Instrument and vice-versa. For example, suppose our training data just capture the sentences (6.1)–(6.5). Then, the Agent noun group contains all nouns that appear as transitive subject, i.e. "husband" (2x), "wife"—and also "hammer". Conversely, the noun group representing Instruments contains all nouns which appear as in a *with*-PP or as a transitive subject in sentences without a *with*-PP, i.e. "hammer" (2x), "chair"—and also "husband". The semantic filters which I will describe in section 6.5 are intended to eliminate noise of such kind, i.e. to fade out the evidence that is erroneously provided by "hammer" regarding the Agent role and by "husband" regarding the Instrument role.[2]

6.2.1 Employing the LSC Model

For the task of creating clusters of frame-argument configurations, it is crucial to find appropriate criteria for clustering. In other words, the problem is to define characteristics of frame-argument configurations which can be employed to decide which of them realise the same roles. In the following, I

[2]Note that the presentation is somewhat simplifying here, in order to make clear the basic idea. In fact, the heuristics proposed in section 6.3 will assign both role types Agent and Instrument to both groups. This is motivated by linguistic facts, which turn out to be more complex than the example discussed here suggests. Nonetheless, the role-discriminating function of semantic filters holds as indicated.

will argue that the latent semantic classes (LSC) model which I introduced in section 3.3 and used already for word sense disambiguation (cf. section 5.2) provides exactly the kind of information needed to define such characteristics. To start this argument, I will recapitulate the rationale behind the LSC approach. The goal of the LSC technique as proposed in (Rooth et al. 1998) is to obtain soft clusters of verb–noun pairs so that similar pairs are gathered in the same cluster. More exactly, the pairs to be clustered consist of a combination of a verb and a frame-argument configuration on the one hand and a noun on the other hand. Thus, in terms of the setting in this thesis, the LSC technique clusters training data items. As this section concentrates on frame-argument configurations, I introduce a formal notation of the pairs to be clustered which emphasises this aspect: For the left component of a data item (like *break#subj:obj/subj*, I will use the expression $fa(v)$ (instead of just v), where fa represents a frame-argument configuration and v a verb. For such a combination of a verb and a frame-argument configuration, I will use the term *verb-frame-argument configuration*.

Formally, an LSC model consists of a set of latent semantic classes where each class c contains $(fa(v), n)$ pairs. These classes are soft clusters, i.e. a pair is member of a cluster *to a certain degree*. This graded membership is modelled in the following way: For each class c and pair $(fa(v), n)$, there is a conditional probability $p(fa(v), n|c)$, the probability that this pair is a member of this class. The crucial idea of the LSC method is that it does not directly estimate this joint probability $p(fa(v), n|c)$. Instead, for each class c, the LSC algorithm estimates the two marginal distributions $p(fa(v)|c)$ and $p(n|c)$. The joint probability is obtained from these marginal probabilities as in equation (6.8):

$$p(fa(v), n|c) = p(fa(v)|c)p(n|c) \tag{6.8}$$

Technically, equation (6.8) formalises the assumption that, given a class c, the probabilities $p(fa(v)|c)$ and $p(n|c)$ are independent from each other. Conceptually, this independence assumption models an abstraction step over data items that constitutes the main characteristic of the LSC model: A class is not just a collection of $(fa(v), n)$ pairs. Rather, it comprises a collection of similar verb-frame-argument configurations on the one hand and a collection of similar nouns on the other hand so that the verb-frame-argument configurations and the nouns which tend to co-occur significantly belong to the same latent semantic class.[3]

Figure 6.2 shows class c_{18} of the LSC model I obtained from the training data (assuming 35 classes and running the EM algorithm for 400 iterations). In particular, the figure displays the verb-frame-argument configurations and nouns with the highest conditional probabilities given c_{18}. At this point, I have to provide two remarks concerning technical subtleties reflected in the notation. Firstly, as for the experiments in chapter 5, I conflated all verb and noun forms which are not captured by WordNet into a single form "notinwordnet". This leads to a significant reduction of parameters, which facilitates processing the model. Secondly, the numbers at the end of the verb-frame-argument configurations indicate the position of the respective argument in the subcategorisation frame. For example, in *put#subj:obj:pp.on/pp.on3* the argument in question, *pp.on*, is the third argument in the associated frame. In most cases, this information is redundant so that I generally do not display the trailing number when displaying verb-frame-argument configurations. However, there are frames where a grammatical function occurs more than once, e.g. double object constructions. In these cases, the position of the argument in question is needed to identify which part of the frame is meant. For example, *give#subj:obj:obj/obj2* represents the first object (which is the second argument in the frame),

[3]Recall that in the context of soft clusters, the expression *belongs to a class* is a simplification which I use for the sake of readability. Exactly, *x belongs to class c* is an abbreviation for *the probability* $p(x|c)$ *of x given class c is (comparably) high*.

$p(c_{18}) = 0.0146$			
v	$p(fa(v)\|c_{18})$	n	$p(n\|c_{18})$
open#subj:obj/obj2	0.0369	eye	0.0976
open#subj/subj1	0.0294	door	0.0776
notinwordnet	0.0253	notinwordnet	0.0385
close#subj:obj/obj2	0.0194	face	0.0328
put#subj:obj:pp.on/pp.on3	0.0123	mind	0.0277
cross#subj:obj/obj2	0.0111	mouth	0.0193
put#subj:obj:pp.in/pp.in3	0.0104	window	0.0173
go#subj:pp.to/pp.to2	0.0098	bed	0.0156
sit#subj:pp.on/pp.on2	0.0094	table	0.0155
hit#subj:obj/obj2	0.0088	chair	0.0143
fill#subj:obj/obj2	0.0088	gray	0.0140
sit#subj:pp.in/pp.in2	0.0073	wall	0.0133
change#subj:obj/obj2	0.0070	side	0.0129
close#subj/subj1	0.0067	line	0.0114
clear#subj:obj/obj2	0.0063	floor	0.0113
...	...	...	...

Figure 6.2: Latent semantic class 18

give#subj:obj:obj/obj3 the second object in a double object construction with "give". For procedural convenience, my implementation always includes the position number in the verb-frame-argument configurations. Therefore, when I present concrete models which I acquired from the data, I mention this technical detail.

Recall how to read figure 6.2: The model defines a probability of membership in class c_{18} for *any possible* $(fa(v), n)$ pair. Of course, pairs consisting of a verb-frame-argument configuration and a noun with high probabilities $p(fa(v)|c_{18})$ and $p(n|c_{18})$, respectively, receive particularly high membership probabilities. Thus, any combination of an item on the left and an item on the right in the figure has a comparably high membership probability. For example, the probability of the pair *(close#subj:obj/obj2, door)* is as follows: $p(\textit{close\#subj:obj/obj2}, door|c_{18}) = p(\textit{close\#subj:obj/obj2}|c_{18}) \times p(door|c_{18}) = 0.0194 \times 0.0776 = 0.0015$.

Like the other LSC examples we have seen so far, this example illustrates that a class tends to comprise similar verb-frame-argument configurations (verbs of opening/closing or putting/moving) as well as similar nouns (parts of faces, parts of buildings, furniture). To understand why these classes provide the information we need to group frame-argument configurations, we have to look more closely at the nature of this similarity, i.e. the question *in what respect* the verb-frame-argument configurations or nouns, respectively, occurring in one class are similar. In section 5.2, I described how words in the same class can mutually disambiguate each other. In particular, I utilised techniques which measure semantic similarity in terms of distances in the WordNet hierarchy. Hence, the disambiguation approach I proposed rests on the assumption that words whose senses are close together in WordNet tend to occur in the same class. The examples of latent semantic classes we have seen so far confirm this behaviour. Furthermore, the comparison of experiments with and without disambiguated data in section 5.5 shows that the WSD approach is effective, which provides further evidence for this

$p(c_{12}) = 0.0183$				
v	$p(fa(v)\|c_{12})$		n	$p(n\|c_{12})$
see#subj:obj:obj/obj3	0.0601		notinwordnet	0.1164
notinwordnet	0.0307		hand	0.0638
wear#subj:obj/obj2	0.0271		arm	0.0252
put#subj:obj:pp.on/obj2	0.0107		light	0.0165
get#subj:obj/obj2	0.0096		foot	0.0164
pick_up#subj:obj/obj2	0.0083		lip	0.0154
remove#subj:obj/obj2	0.0083		face	0.0153
break#subj:obj/obj2	0.0081		hair	0.0134
place#subj:obj:pp.on/obj2	0.0078		finger	0.0131
touch#subj:obj/obj2	0.0077		shoulder	0.0113
begin#subj:to/subj1	0.0064		heart	0.0108
find#subj:obj/obj2	0.0058		body	0.0099
cover#subj:obj:pp.with/pp.with3	0.0058		glass	0.0095
hold#subj:obj/obj2	0.0054		leg	0.0093
light#subj:obj/obj2	0.0052		back	0.0077
...	...		...	...

Figure 6.3: Latent semantic class 12

assumption.

However, this captures only one aspect of the similarity of the items clustered together in an LSC model. The primary nature of this similarity is more sophisticated: Those verb-frame-argument configurations that behave similarly *w.r.t. the nouns with which they co-occur* are gathered in the same class(es), and, conversely, those nouns that behave similarly *w.r.t. the verb-frame-argument configurations with which they co-occur* are gathered in the same class(es). For instance, class c_{18} models the finding that verbs like "open", "close", "put", "go", or "hit" tend to select nouns like "eye", "door", "mouth", "window", or "bed" as arguments. Essentially, latent semantic classes sort verb-frame-argument configurations and nouns according to their *mutual relational similarity*.

In other words, a latent semantic class as a whole represents a certain type of semantic relationship between verb-frame-argument configurations and nouns. This semantic relationship has two aspects, which may be more or less distinct in an individual class. As yet, we have concentrated on one of these aspects: verbs and nouns in a certain class tend to be associated with one or a couple of certain semantic fields. I make use of this tendency by employing the LSC model for word sense disambiguation. In the context of linking, however, the other aspect is crucial: Actually, it turns out that a number of classes strongly indicate a certain thematic role. More precisely, it is often the case that the relations between verb-frame-argument configurations and nouns in a certain class tend to correspond to the same thematic role. In the following, I will discuss several classes where both aspects are visible.

Figures 6.3 and 6.4 show the classes c_{12} and c_{26}, respectively. As can easily be seen, both classes model verb–noun relations which realise the Patient role. However, they differ in the semantic fields they cover. Class c_{12} comprises relations between mainly verbs of placing or displacing and their

$p(c_{26}) = 0.0303$			
v	$p(fa(v)\|c_{26})$	n	$p(n\|c_{26})$
seem#subj:to/subj1	0.0215	notinwordnet	0.1257
notinwordnet	0.0169	people	0.0651
kill#subj:obj/obj2	0.0132	man	0.0391
die#subj/subj1	0.0132	child	0.0363
become#subj:obj/obj2	0.0117	woman	0.0275
begin#subj:to/subj1	0.0099	patient	0.0166
tend#subj:to/subj1	0.0095	member	0.0105
appear#subj:to/subj1	0.0087	student	0.0100
give#subj:obj:obj/obj2	0.0069	one	0.0096
become#subj:ap/subj1	0.0068	girl	0.0096
allow#subj:obj:to/obj2	0.0067	family	0.0093
give#subj:obj:pp.to/pp.to3	0.0063	person	0.0092
come#subj:to/subj1	0.0060	mother	0.0075
see#subj:obj/obj2	0.0059	boy	0.0069
look#subj:pp.like/pp.like2	0.0058	animal	0.0066
...	...	...	...

Figure 6.4: Latent semantic class 26

objects, mostly body parts. In contrast, class c_{26} captures relations between verbs of diverse semantic domains and nouns mainly denoting persons.

In contrast, figures 6.5–6.7 show three classes, c_3, c_{17}, and c_{30}, which model verb–noun relations corresponding to the Agent role. Again, these classes differ in the semantic fields to which the verbs and nouns belong. c_3 apparently does not exhibit a particular semantic domain, except from the fact that the nouns denote persons or person groups. c_{17} focuses on actions of exerting force, where the agentive noun arguments may be human or inanimate. c_{30} captures actions of communication or cognition, naturally with human arguments.

Looking at concrete verbs and their distributions within one class or across different classes shows how the LSC model can be utilised for grouping frame-argument configurations. First of all, there are verbs which occur in a class with several frame-argument configurations. For example, class 26 contains two different verb-frame-argument configurations pertaining to "give": *give#subj:obj:obj/obj2* and *give#subj:obj:pp.to/pp.to3*. This is an instance of the dative alternation (which I will discuss in section 6.3.2). Both frame-argument configurations realise the same semantic argument for "give" (the recipient). This illustrates that an LSC model is able to assemble different syntactic realisations of the same semantic argument of a verb in the same class. I mentioned this property of the LSC method already in section 3.3.

On the other hand, the same verb-frame-argument configuration may occur in different classes. For example, *take#subj:obj/subj1* and *tell#subj:obj/subj1* have high probabilities in c_3 and c_{30}, *give#subj:obj:obj/subj1* has high probabilities in c_3 and c_{17}. Note that the occurrence of a verb-frame-argument configuration in different classes does not necessarily coincide with different meanings of the respective verb. There is no apparent meaning difference for "tell" in the context of c_3 and c_{30},

$p(c_3) = 0.0163$				
v	$p(fa(v)\|c_3)$		n	$p(n\|c_3)$
make#subj:obj/subj1	0.0558		notinwordnet	0.4306
take#subj:obj/subj1	0.0543		he	0.1060
win#subj:obj/subj1	0.0303		who	0.0349
tell#subj:obj/subj1	0.0218		party	0.0084
give#subj:obj/subj1	0.0209		father	0.0080
play#subj:obj/subj1	0.0190		man	0.0073
hold#subj:obj/subj1	0.0146		company	0.0059
become#subj:obj/subj1	0.0145		team	0.0054
give#subj:obj:obj/subj1	0.0126		family	0.0042
notinwordnet	0.0121		labour	0.0042
lose#subj:obj/subj1	0.0089		mother	0.0041
run#subj:obj/subj1	0.0085		group	0.0040
join#subj:obj/subj1	0.0084		woman	0.0036
own#subj:obj/subj1	0.0078		government	0.0031
tell#subj:obj:that/subj1	0.0077		kingbolt	0.0030
...	...		...	...

Figure 6.5: Latent semantic class 3

$p(c_{17}) = 0.008$				
v	$p(fa(v)\|c_{17})$		n	$p(n\|c_{17})$
hit#subj:obj/subj1	0.0211		notinwordnet	0.1171
kill#subj:obj/subj1	0.0195		police	0.0356
catch#subj:obj/obj2	0.0190		force	0.0234
notinwordnet	0.0134		car	0.0168
carry#subj:obj/subj1	0.0119		man	0.0147
give#subj:obj:obj/subj1	0.0117		firebreak	0.0143
begin#subj:to/subj1	0.0114		troop	0.0140
strike#subj:obj/subj1	0.0109		bomb	0.0120
come#subj/subj1	0.0099		army	0.0113
destroy#subj:obj/subj1	0.0098		train	0.0099
attack#subj:obj/subj1	0.0096		someone	0.0092
arrive#subj/subj1	0.0095		rain	0.0091
stop#subj/subj1	0.0091		sight	0.0074
reach#subj:obj/subj1	0.0086		light	0.0071
leave#subj:obj/subj1	0.0084		shot	0.0071
...	...		...	...

Figure 6.6: Latent semantic class 17

$p(c_{30}) = 0.1081$			
v	$p(fa(v)\|c_{30})$	n	$p(n\|c_{30})$
say#subj/subj1	0.1129	notinwordnet	0.5173
say#subj:adv/subj1	0.0173	he	0.4198
notinwordnet	0.0172	man	0.0074
tell#subj:obj/subj1	0.0117	woman	0.0032
ask#subj/subj1	0.0112	doctor	0.0025
think#subj/subj1	0.0087	mother	0.0024
add#subj/subj1	0.0087	girl	0.0023
try#subj:to/subj1	0.0076	father	0.0022
make#subj:obj/subj1	0.0075	boy	0.0021
want#subj:to/subj1	0.0069	someone	0.0015
will#subj:vbase/subj1	0.0068	who	0.0012
begin#subj:to/subj1	0.0066	one	0.0010
take#subj:obj/subj1	0.0059	voice	0.0009
write#subj/subj1	0.0055	husband	0.0006
go#subj:to/subj1	0.0054	people	0.0006
...	...	...	...

Figure 6.7: Latent semantic class 30

respectively. Rather, since "tell" fits into the semantic domains which both classes focus on, and *tell#subj:obj/subj1* expresses a semantic role which is expressed in both classes (Agent), this verb-frame-argument configuration unsurprisingly occurs in these two classes. In general, the role which corresponds to a verb-frame-argument configuration is not reflected by its membership in one particular class, but by its *overall profile of membership* in all classes of the LSC model. This membership profile depends on both the individual verb (it differs for "give" and "tell") and the role that is expressed by the frame-argument configuration associated with that verb (it differs for the subject and the object of "give"). The approach I propose for grouping frame-argument configurations assumes that different frame-argument configurations pertaining to the same verb and expressing the same role (e.g. *give#subj:obj:obj/obj2* and *give#subj:obj:pp.to/pp.to3*) have similar class membership profiles.

As the LSC model represents the degreed class membership by probabilities, it is straightforward to model the membership profile just discussed by probabilities as well. Actually, the class membership profile of a verb-frame-argument configuration $fa(v)$ can be straightforwardly expressed by a single probability distribution, namely the distribution of classes c_i given that configuration, i.e. $p(c_i|fa(v))$. Intuitively, this probability distribution models how $fa(v)$ is divided among the classes in the LSC model. The LSC model does not immediately estimate these probabilities. Its parameters comprise marginal class probabilities $p(c_i)$ and class membership probabilities $p(fa(v)|c_i)$. However, with these probability distributions, the probabilities $p(c_i|fa(v))$ can be retrieved using Bayes' law:

$$p(c_i|fa(v)) = \frac{p(c_i, fa(v))}{p(fa(v))} = \frac{p(fa(v)|c_i)p(c_i)}{\sum_{c_j} p(fa(v)|c_j)p(c_j)} \tag{6.9}$$

Equation (6.9) makes explicit that $p(c_i|fa(v))$ is a function of the class membership probability

$p(fa(v)|c_i)$ and thus the distribution obtained by this formula captures the information the LSC model provides about the distributional profile of $fa(v)$. The fact that this profile can be summarised by a single probability distribution is very convenient concerning the task addressed in this section, i.e. grouping frame-argument configurations. According to the previous paragraphs, this task amounts to dividing verb-frame-argument configurations $fa_1(v), \ldots, fa_n(v)$ pertaining to a particular verb v into groups so that configurations which have a similar distributional profile are grouped together. As the distributional profile indicates the thematic role that a verb-frame-argument configuration realises, it is likely that a group of several $fa_k(v)$ with similar profiles comprises frame-argument configurations which express the same role. Of course, a precondition for this approach is that we are able to measure the similarity of distributional profiles corresponding to different verb-frame-argument configurations. As these profiles are represented as single probability distributions, this can be done in a simple way. Several formulae for measuring the (dis-)similarity of probability distributions have been proposed in the literature. Thus, the similarity of two verb-frame-argument configurations $fa_k(v)$ and $fa_l(v)$ can be measured by computing the similarity of the corresponding probability distributions $p(c_i|fa_k(v))$ and $p(c_i|fa_l(v))$.

(Manning & Schütze 1999, p. 303–306) review some measures of the distance between two probability distributions proposed in the literature (following (Dagan, Lee & Pereira 1997)). In section 3.4.1.2, we have already encountered a very common information-theoretic distance measure: relative entropy. The *relative entropy* (or *Kullback-Leibler distance*) $D(p\|q)$ of two distributions p and q is defined as in equation (6.10) (which is a repetition of equation (3.19) on page 67):

$$D(p\|q) = \sum_x p(x) \log \frac{p(x)}{q(x)} = \sum_x p(x)(\log \frac{1}{q(x)} - \log \frac{1}{p(x)}) \tag{6.10}$$

Relative entropy is a measure of the information that is lost if one assumes distribution q while the true distribution is p. The main problem of this measure for our purposes is that it is not symmetric, i.e. in general, $D(p\|q) \neq D(q\|p)$. Thus, if we want to assess the similarity of two configurations $fa_k(v)$ and $fa_l(v)$, it is not clear whether we should use $D(p(c_i|fa_k(v))\|p(c_i|fa_l(v)))$ or $D(p(c_i|fa_l(v))\|p(c_i|fa_k(v)))$. The problem is that, in our setting, we do not have to compare two probability distributions where one of them is "true" and the other one is a more or less good approximation of it. Instead, we have to compare two distributions of equal status, since each of them represents the distributional profile of a certain verb-frame-argument configuration. Therefore, we need a symmetric (dis-)similarity measure.[4] A distance measure which is derived from relative entropy is *information radius* ($IRad$), which is defined as follows:

$$IRad(p,q) = D(p\|\frac{p+q}{2}) + D(q\|\frac{p+q}{2}) \tag{6.11}$$

The information radius measures how much information is lost if we model two random variables whose underlying probability distributions are p and q, respectively, by the average distribution $\frac{p+q}{2}$. In our setting, this means the following: We have to compare two verb-frame-argument configurations $fa_k(v)$ and $fa_l(v)$ whose profiles are modelled by the distributions $p(c_i|fa_k(v))$ and $p(c_i|fa_l(v))$, respectively. The information radius of these two distributions $IRad(p(c_i|fa_k(v)), p(c_i|fa_l(v)))$ measures the information which would be lost if we modelled the profiles of both configurations by the

[4]The other main drawback of relative entropy is that this measure is not defined for those x for which $q(x) = 0$ and $p(x) \neq 0$. However, this is no problem in our context. The parameters estimated by the LSC algorithm are non-zero, i.e. $p(c_i) > 0$ and $p(fa(v)|c_i) > 0$ for all c_i and $fa(v)$. Thus, following equation (6.9), $p(c_i|fa(v)) > 0$ for all c_i and $fa(v)$.

average of these two distributions, i.e. using the following distributions $\hat{p}(...)$ instead of the "true" ones:

$$\hat{p}(c_i|fa_k(v)) = \hat{p}(c_i|fa_l(v) = \frac{p(c_i|fa_k(v)) + p(c_i|fa_l(v))}{2} \tag{6.12}$$

The information radius is a symmetric measure, i.e. $IRad(p,q) = IRad(q,p)$. Thus, it is appropriate for our purpose. It can be shown that this measure has a fixed range: $0 \leq IRad(p,q) \leq 2\log 2$ for all distributions p and q.

(Manning & Schütze 1999) discuss a third distance measure, the L_1 *norm*, which is defined as $\sum_x |p(x) - q(x)|$. This measure is also symmetric, i.e. adequate for our needs. I decided to use $IRad$ for my work. This is a purely pragmatic choice, motivated by experiments reported in (Dagan et al. 1997) (cited by (Manning & Schütze 1999)) and (Dagan, Lee & Pereira 1999), which show that $IRad$ performs better than L_1 in several tasks. However, there is no principled reason against using the L_1 norm for the grouping strategy I describe below.

Employing $IRad$ to measure the distance $dist$ between two verb-frame-argument configurations $fa_k(v)$ and $fa_l(v)$ leads to the following equation:

$$dist(fa_k(v), fa_l(v)) = IRad(p(c_i|fa_k(v)), p(c_i|fa_l(v))) \tag{6.13}$$

I made use of this distance measure in a straightforward way: If $dist(fa_k(v), fa_l(v))$ does not exceed a certain similarity threshold thr_{sim}, then $fa_k(v)$ and $fa_l(v)$ are put into the same group. I will explain below how I determine this similarity threshold.

Now we have the means required to group frame-argument configurations of a certain verb according to their similarity w.r.t. the semantic roles they tend to realise. The general strategy I pursue for this task is as follows:

1. For each verb v, let $fa_1(v), \ldots, fa_n(v)$ be the verb-argument configurations pertaining to v which occur in the data
2. create a set of groups $G = \{g_1, \ldots, g_m\}$ of these verb-frame-argument configurations so that similar configurations are pooled in the same group. In particular, the following conditions must hold:
   - **correctness:** all configurations in a group must be pairwise similar: if $fa_k(v) \in g_i$ and $fa_l(v) \in g_i$, then $dist(fa_k(v), fa_l(v)) < thr_{sim}$, (i.e. if two configurations are in the same group, then the distance between them must not exceed the similarity threshold)
   - **completeness:** all similar configurations must be grouped together: if $dist(fa_k(v), fa_l(v)) < thr_{sim}$, then there must be a group g_i so that $fa_k(v) \in g_i$ and $fa_l(v) \in g_i$ (i.e. if two configurations are similar, then there must be a group which contains both configurations)
   - **non-redundancy:** the set of groups G must be non-redundant: if $g_i, g_j \in G$, then $g_i \not\subseteq g_j$ and $g_j \not\subseteq g_i$ (i.e. there must not be a group that is a subset of another group)

- **exhaustive coverage:** each configuration $fa_k(v)$ must be member of at least one group $g_i \in G$. If $fa_k(v)$ is not similar to any other configuration pertaining to v, then it is captured by a separate group containing solely this configuration

To retrieve this set of groups, I use a simple inductive approach: I start with the first verb-frame-argument configuration $fa_1(v)$ (the order in which the argument configurations are processed is completely arbitrary) and create a group containing just this configuration. Each of the subsequent inductive steps processes a further verb-frame-argument configuration $fa_k(v)$. In particular, it is examined whether this configuration can be integrated into one or more of the existing groups. To achieve this, $fa_k(v)$ is compared to all members of each group. If $fa_k(v)$ is pairwise similar to all members of a group g_i, then it is integrated into that group. If $fa_k(v)$ is only similar to *some* members of a group g_i (but not to all), then a new group is created that contains $fa_k(v)$ and those configurations in g_i that are similar to it.[5] If $fa_k(v)$ is not similar to any other frame-argument configuration, then a separate group containing just this configuration is created. After all frame-argument configurations pertaining to the same verb have been processed, a final step checks the resulting groups for subsets: the groups are compared pairwise; if a group is a subset of another group, then it is deleted.

Let us look at an example that illustrates this approach. Recall the verb-frame-argument configurations pertaining to "break" mentioned above:

- *break#subj/subj*
- *break#subj:obj/subj*
- *break#subj:obj/obj*
- *break#subj:obj:pp,with/subj*
- *break#subj:obj:pp,with/obj*
- *break#subj:obj:pp,with/pp.with*

In the following, I will demonstrate how the inductive strategy retrieves the frame-argument configuration groups for that verb which I discussed above. The approach starts by creating a group g_1 that contains the first configuration *break#subj/subj*:

$g_1 = \{$*break#subj/subj*$\}$

The next step processes the next configuration, i.e. *break#subj:obj/subj*. This configuration is compared with the only member of the only existing group g_1 (*break#subj/subj*) They are found not to be similar. Thus, a new group g_2 is created to capture *break#subj:obj/subj*:

$g_1 = \{$*break#subj/subj*$\}$
$g_2 = \{$*break#subj:obj/subj*$\}$

In the next step, *break#subj:obj/obj* is processed. This configuration is compared with the only member of g_1, *break#subj/subj*. As the two configurations are similar, *break#subj:obj/obj* is integrated

[5] Note that the mutual similarity of these configurations has been verified already, since they are members of the same group g_i.

into g_1. The comparison of *break#subj:obj/obj* with the member of g_2, *break#subj:obj/subj*, yields no similarity. Thus, g_2 remains unchanged. Altogether, the groups after that step look like this:

$g_1 = \{$*break#subj/subj*, *break#subj:obj/obj*$\}$
$g_2 = \{$*break#subj:obj/subj*$\}$

The next step processes the configuration *break#subj:obj:pp,with/subj*. The comparison with the configurations in the two groups yields that it is not similar to any member of g_1, but similar to the only member of g_2. Thus, *break#subj:obj:pp,with/subj* is integrated into g_2, while g_1 remains unchanged:

$g_1 = \{$*break#subj/subj*, *break#subj:obj/obj*$\}$
$g_2 = \{$*break#subj:obj/subj*, *break#subj:obj:pp,with/subj*$\}$

The next step deals with *break#subj:obj:pp,with/obj*. This configuration is similar to all members of g_1 and to no member of g_2 so that it is integrated in g_1, leaving g_2 unchanged:

$g_1 = \{$*break#subj/subj*, *break#subj:obj/obj*, *break#subj:obj:pp,with/obj*$\}$
$g_2 = \{$*break#subj:obj/subj*, *break#subj:obj:pp,with/subj*$\}$

In the last inductive step, *break#subj:obj:pp,with/pp.with* is processed. This configuration is not similar to any configuration in g_1. Regarding g_2, we have the slightly more complex case that *break#subj:obj:pp,with/pp.with* is not similar to all members, but only to one of them, namely *break#subj:obj/subj*. In this case, a new group g_3 has to be created that comprises the new configuration *break#subj:obj:pp,with/pp.with* and the similar configuration of g_2. Hence, the groups resulting from these induction steps are:

$g_1 = \{$*break#subj/subj*, *break#subj:obj/obj*, *break#subj:obj:pp,with/obj*$\}$
$g_2 = \{$*break#subj:obj/subj*, *break#subj:obj:pp,with/subj*$\}$
$g_3 = \{$*break#subj:obj/subj*, *break#subj:obj:pp,with/pp.with*$\}$

The final check for subsets yields that none of these groups is a subset of another one so that none of them has to be deleted.

Looking at these groups (which are identical to the groups presented at the beginning of section 6.2), it is easy to see that they correspond to different thematic roles: g_1 corresponds to the Patient, g_2 to the Agent, and g_3 to the Instrument. Furthermore, this example illustrates that a verb-frame-argument configuration may be a member of more than one group. In this case, *break#subj:obj/subj* is in g_2 and in g_3. This reflects the observation discussed above that this configuration might express an Agent or an Instrument.

For the sake of better readability, I will henceforth use a more concise notation for argument groups. Technically, such groups contain verb-frame-argument configurations. However, since all configurations in a group pertain to the same verb, I will mention this verb as a parameter of the group identifier and denote the group members by the bare frame-argument configurations. For example, the group g_3 above is written as

$g_3(break) = \{$*subj:obj/subj*, *subj:obj:pp,with/pp.with*$\}$

One issue which I have not addressed so far is how to fix the similarity threshold that determines whether two verb-frame-argument configurations are similar or not. Several possibilities to define a threshold are conceivable. As noted above, the information radius, which I employ as distance

measure, ranges from 0 (identity of the compared distributions) to $2 \log 2$ (maximal distance). Any value between these two limits could serve as a threshold. Furthermore, one could either use a global threshold, i.e. the same threshold for all verbs, or use different thresholds for the frame-argument configurations pertaining to different verbs. For simplicity, I decided to use a global threshold and to determine its value empirically. More precisely, I employ the mean of all distances which are calculated for the grouping task as the similarity threshold. This mean is obtained by a preprocessing step: For all verbs v, the distance between each possible pair of frame-argument configurations pertaining to v is computed, and the arithmetic mean of all these distances is calculated. This arithmetic mean is adopted as the threshold: If a distance between two verb-frame-argument configuration is below this value, then these configurations are considered to be similar.

6.2.2 A Linguistic Interpretation of the LSC Model

Before I report results of applying the approach described above, I would like to outline a possible linguistic interpretation of this approach and the LSC model in general. (Rooth et al. 1998) themselves offer a linguistic interpretation of their approach in the context of theories of lexical semantic representations. In short, they suggest that the labels of LSC classes c_i should be used as types of verb–argument relations in semantic representations of verbs (cf. (Rooth et al. 1998, p. 112–116)). In contrast, the interpretation I will sketch in this section aims to relate my grouping approach to a linguistic theory of thematic roles.

In general, a latent semantic class models a certain semantic relationship between verbs and nouns. We have seen various classes where this relationship corresponds to a specific thematic role (often connected with a certain semantic domain). Unsurprisingly, not all classes exhibit a neat correspondence with a single thematic role. Often the kind of semantic relation captured by a class is not immediately obvious. However, in some cases such a relation resembles another theoretic construct related to thematic roles, which I described in section 2.1.3: a proto role entailment à la Dowty. For example, class c_{18} (displayed in figure 6.2 on page 157) cannot be associated with a specific role. The verb-frame-argument configurations express different roles (e.g. *put#subj:pp.on/pp.on2* a Goal, *sit#subj:pp.on/pp.on2* a Location, or *hit#subj:obj/obj2* a Patient). However, the verb–noun relations represented in that class commonly imply that the Proto-Patient entailment "stationary relative to movement of another participant" holds for the noun arguments.

The finding that latent semantic classes may correspond to proto role entailments leads to the linguistic interpretation I suggest here. Actually, there is a striking analogy between an LSC model and Dowty's theory of thematic roles. This theory states that roles are no discrete categories, but prototypical concepts, and that arguments exhibit a degreed association to different role types. In particular, the proto roles P-Agent and P-Patient are each characterised by a set of proto entailments. In a certain sentence, the argument that meets the most P-Agent entailments is the P-Agent, and the argument that meets the most P-Patient entailments is the P-Patient. Thus, an argument that represents a certain role does not have to meet all entailments corresponding to that role. Rather, the fulfilment or non-fulfilment of an entailment by an argument provides evidence for or against the possibility that this argument represents the role corresponding to that entailment. The cumulative evidence regarding all entailments determines the role which an argument expresses. Moreover, Dowty explicitly leaves open the possibility that the individual entailments have different weights so that an entailment with a higher weight provides more evidence for the corresponding role than an entailment with a lower weight.

These considerations lead to the idea that, in the broadest sense, an LSC model can be regarded as a probabilistic implementation of Dowty's theory: We have seen that latent semantic classes can be viewed as corresponding to proto role entailments (or entire roles). Therefore, analogously to Dowty's approach where the fulfilment of entailments provides cumulative evidence for the role of an argument, one can say that the profile of membership in latent semantic classes provides cumulative evidence for the role which an argument expresses. (Here, *argument* means a frame-argument configuration, not an individual complement in a particular sentence.[6]) In the following, I explicate this analogy in more detail.

Basically, the collection of classes in the LSC model corresponds to the collection of proto role entailments in Dowty's theory. Within these collections, the function of an individual class mirrors the function of an individual entailment. This does not mean that the verb–noun relation which is captured by a certain class can always be interpreted in terms of an entailment like those proposed by Dowty.[7] Some classes correspond to such entailments, some classes correspond to entire roles, and for some classes, there is no straightforward interpretation related to thematic roles at all. However, like proto role entailments, latent semantic classes can be viewed as properties which, in total, provide the information that can be utilised to determine the role expressed by a certain frame-argument configuration of a certain verb. In Dowty's approach, this information comprises the pattern of fulfilment/non-fulfilment of the individual proto entailments. In the LSC approach, this information is the degreed membership profile of the respective verb-frame-argument configuration $fa(v)$. As described in the previous subsection, this profile is modelled by the distribution $p(c_i|fa(v))$. This distribution is computed by equation (6.9), repeated here:

$$p(c_i|fa(v)) = \frac{p(fa(v)|c_i)p(c_i)}{\sum_{c_j} p(fa(v)|c_j)p(c_j)} \tag{6.14}$$

In its basic form, Dowty's theory is discrete in the sense that an entailment is either fulfilled or not fulfilled by the argument in question. In contrast, the LSC approach is statistical; instead of the binary fulfilment vs. non-fulfilment classification it estimates a probability, namely $p(c_i|fa(v))$. However, Dowty explicitly acknowledges that the kind of entailments he assumes do not have entirely clear-cut boundaries and that they may hold to differing degrees, depending on the situation and their participants denoted in the sentence in question. The work I present in this thesis is based on the assumption that probabilistic models are suitable to account just for such degreed phenomena. Moreover, as noted above, Dowty states that it might be more appropriate to associate different weightings to different entailments. He considers both *a priori* weightings of entailments regardless of a particular verb (cf. (Dowty 1991, p. 574)) and weightings which depend on the examined verb or verb class (cf. (Dowty 1991, p. 597)). Both kinds of weighting are reflected in the numerator in equation (6.14): $p(c_i)$ mirrors an *a priori* weight, $p(fa(v)|c_i)$ a verb-dependent weight.

The major structural difference between Dowty's theory and the argument grouping approach I propose here is the way in which the profiles of entailment fulfilment or class membership, respectively,

[6] In Dowty's theory, the fulfilment of an entailment by an argument must follow from the verb's meaning, and not be a consequence of other coincidental factors, e.g. the meaning of a complement or the context. For example, for "kill", the subject does not meet the P-Agent entailment of volitional involvement, nor the entailment of sentience (an illness also can kill someone), although it is straightforward to think of concrete sentences where the subject is a volitional and sentient actor. However, for "murder", these entailments are met by the subject—this is implied by the meaning of the verb. Thus, entailments are meant to be valid or not valid for frame-argument configurations, irrespective of individual arguments in concrete sentences.

[7] However, note that Dowty does not claim that the catalogue of entailments he proposes is exhaustive.

are employed within the overall linking approach. This difference is connected with the differing ultimate goals of the two approaches. Like all linguistic linking theories, Dowty's approach aims at explaining the correspondence between syntactic arguments and their underlying thematic roles in an individual sentence. Therefore, the entailment profiles of all arguments in this sentence are compared to each other; the realisation of the Agent as subject and the Patient as object is explained by the fact that the former argument meets the most P-Agent entailments and the latter one meets the most P-Patient entailments. In contrast to this explanative theory, the linking strategy described in this chapter has the constructive function of sorting data items according to their underlying roles so that overall statistics about the semantic preferences for these roles can be collected. Specifically, the argument grouping approach does not compare the class membership profiles of the arguments in an individual sentence, but the overall membership profiles of different frame-argument configurations of a verb regarding the complete training corpus. Another difference is that this stage of the linking strategy does not have access to information about which class corresponds to which role or role entailment. Frame-argument configurations with similar profiles are collected in one group, since it is expected that these similar profiles indicate the same thematic role. However, the question *which* role is indicated by these profiles is not addressed here. This is determined in the second stage of the linking strategy, which I will describe in section 6.3. In contrast, in Dowty's theory each entailment is explicitly associated with one particular role (Agent or Patient) and the fulfilment of an entailment is understood to provide evidence for the associated role.

In summary, Dowty's proto entailment approach and the LSC approach exhibit intriguing analogies concerning their design, but differ significantly w.r.t. their function within the complete theoretical or technical apparatus they are part of. Therefore, interpreting the LSC approach in terms of Dowty's approach is manifest, but not compulsory.

6.2.3 Experiments

To test the argument grouping method proposed here, I applied this approach as described in section 6.2.1 to the LSC model I estimated from the training data.[8] A manual inspection of the results shows that this approach is effective. However, this inspection also reveals shortcomings which motivate a simple refinement of the approach. To give a general impression of the acquired results, I show some argument groups retrieved in the experiment which exemplify the overall suitability as well as the limitations of the strategy. I start with the verb that I use to illustrate the different stages of my linking approach in this chapter, namely "break". The following groups have been acquired for that verb:

```
break: subj:obj/obj2, subj/subj1

break: subj:pp.into/subj1, subj:obj/subj1, subj/subj1

break: subj:pp.into/pp.into2, subj/subj1
```

Interestingly, this result differs from the illustrative (and artificial) examples for "break" in this chapter (which I nonetheless will continue to use due to their clarity). First of all, none of the groups contains a *with*-PP. This PP occurs as argument of "break" at such a low frequency that it is discarded when

[8] The similarity threshold, i.e. the mean of all distances of pairwise compared frame-argument configurations, was computed as 1.37.

training the LSC model. In exchange, an *into*-PP co-occurs with "break" to a significant extent (as part of the *subj:pp.into* frame). The first group appropriately captures the causative-inchoative alternation (the object of the transitive variant corresponds with the subject of the intransitive variant). The second group assembles the subjects of all argument frames co-occurring with "break" in the LSC model. The third group contains the *into*-PP and the intransitive subject. Intuitively, the first group embodies the Patient, the second one the Agent, and the third one the Goal. However, assuming this role assignment it is linguistically inadequate that the intransitive subject belongs to the second and the third group, because this frame-argument configuration neither realises an Agent (except for elliptic constructions) nor a Goal, but a Patient of "break". Apparently, the reason for including *break#subj/subj1* in all groups is that virtually any kind of entity might break. Hence, the corresponding slot is filled, among others, by nouns which denote typical Agents (e.g. "doctor" "Macedonian") or Goals (e.g. "box", "country", "window"). Therefore, it is not surprising that this verb-frame-argument configuration is classified as similar to those configurations expressing the Agent or the Goal, respectively. This is an example of a general problem of my linking strategy: frame-argument configurations realising different roles may nevertheless be filled with similar nouns so that they are grouped together.

For other verbs, the acquired groups completely conform to linguistic intuition. For example, in many cases the approach is able to recognise whether verbs do or do not undergo the causative/inchoative alternation and retrieve appropriate groups. For example, the groups for "decrease" reflect this alternation:

```
decrease: subj:obj/subj1

decrease: subj:obj/obj2, subj/subj1
```

In contrast, the groups for "eat" indicate that the intransitive variant emerges from dropping the object of the transitive variant so that the subjects of the two variants are correspondent (an additional intransitive frame containing an adverb is also classified correctly):

```
eat: subj:obj/subj1, subj/subj1, subj:adv/subj1

eat: subj:obj/obj2
```

Typically, the groups yielded for a given verb capture the alternations which that verb undergoes, but also contain some noise. For example, "load" receives, among others, the following groups:

```
load: subj:obj:pp.with/subj1, subj:obj:pp.onto/subj1,
      subj:obj:pp.into/subj1, subj:obj/subj1

load: subj:obj:pp.with/obj2, subj:obj:pp.into/pp.into3,
      subj:obj/obj2

load: subj:obj:pp.with/pp.with3, subj:obj:pp.into/obj2,
      subj:obj/obj2, subj/subj1

load: subj:obj:pp.onto/pp.onto3, subj:obj:pp.onto/obj2,
```

```
subj:obj:pp.into/pp.into3, subj:obj:pp.into/obj2,
subj:obj/obj2
```

The first group gathers the subjects of all frames. The second and the third group reflect the spray/load alternation (cf. examples (2.20) and (2.21) on page 22): they appropriately model the fact that the object in a frame containing an *into*-PP corresponds to a *with*-PP and, conversely, the object in a frame containing a *with*-PP corresponds to an *into*-PP. Thus, for the alternation

(6.15)

a. Bill loaded the books into the container.

b. Bill loaded the container with the books.

the second group represents "the container", while the third one represents "the books". The fourth group is partly inadequate. It correctly encodes the correspondence between an *into*-PP and an *onto*-PP. However, it also says that these PPs correspond with the direct object *in their own frames*. This would mean that two syntactic arguments in a sentence represent the same underlying semantic argument, which is obviously inaccurate. Overall, these groups are noisy, but not completely arbitrary. It is not the case that the correct combinations of frame-argument configurations arise by chance, among a lot of other deliberate wrong combinations. Rather, the grouping tends to be adequate, with a certain amount of errors. This example also shows that the groups might miss some correspondences. Here, the groups do not capture the observation that the *onto*-PP undergoes the same alternation with an object as the *into*-PP.

The groups acquired for "fill" demonstrate that the approach is able to capture sophisticated alternation patterns as well as the fact that one frame-argument configuration possibly can express several kinds of semantic arguments:

```
fill: subj:obj:pp.with/subj1, subj:obj/subj1

fill: subj:pp.with/subj1, subj:obj:pp.with/obj2, subj:obj/obj2

fill: subj:pp.with/pp.with2, subj:obj:pp.with/pp.with3,
      subj:obj/subj1

fill: subj:obj:pp.with/obj2, subj:obj/subj1
```

This verb undergoes several overlapping alternations, illustrated below:

(6.16)

a. John filled the tank with water.

b. John filled the tank.

c. Water filled the tank.

d. The tank filled with water.

Concerning these sentences, the first group represents “John”, the second group “the tank”, and the third group “water”. These three groups form a correct and complete model of the alternations shown in (6.16). In particular, the frame-argument configuration *subj:obj/subj1* is instantiated by “John” and “water”. Consequently, this type is member of the first and the third group. The fourth group is erroneous.

Sometimes the approach fails to assemble corresponding frame-argument configurations in a single group, but acquires multiple groups instead (which often overlap). For example, the arguments expressing the Location of “kneel” are collected in two groups:

```
kneel: subj:pp.by/pp.by2, subj:pp.beside/pp.beside2

kneel: subj:pp.on/pp.on2, subj:pp.in/pp.in2, subj:pp.by/pp.by2
```

As we have seen already in the “break” example, it often happens that frame-argument configurations which realise different roles are grouped together because they are semantically similar. The most obvious cases are comitatives, which are usually expressed by a *with*-PP, but are semantically equivalent to the Agent of the respective sentence. Hence, the subject and the *with*-PP are semantically similar and included in one group, e.g. for “cooperate”:

```
cooperate: subj:pp.with/pp.with2, subj:pp.with/subj1,
           subj:pp.in/subj1
```

Source and Goal are often semantically similar as well so that the corresponding PPs are grouped together, as for “rise”:

```
rise: subj:pp.to/pp.to2, subj:pp.from:pp.to/pp.to3,
      subj:pp.from:pp.to/pp.from2
```

On the other hand, many groups are inappropriately acquired although their members apparently do not behave similarly w.r.t. the nouns they take as fillers. For example, “mark” receives a couple of groups like

```
mark: subj:obj:pp.in/pp.in3, subj:obj:pp.in/obj2, subj:obj/obj2,
      subj:obj/subj1
```

where subjects, objects, and PPs are included in one group. Often (as in this case) the reason is that the nouns that fill these frame-argument configurations exhibit a high semantic variety, denoting humans, inanimate entities, actions, or abstract notions. It appears that this variety leads to comparably similar class distributions of the different frame-argument configurations so that the computed distance between them is low.

Overall, the results presented here are promising. They suggest that the argument grouping strategy is suitable for our needs, despite several weaknesses of the kind that any automatic clustering method

exhibits. One of these shortcomings can easily be eliminated. We have seen examples of groups which include frame-argument configurations with the same frame (e.g. *subj:obj/obj2* and *subj:obj/subj1*). As noted, this is inappropriate, since a semantic argument is usually not expressed by multiple syntactic arguments at a time. Therefore, I modified the grouping algorithm by adding the constraint that two frame-argument configurations with the same frame must not be grouped together.

6.3 Heuristics to Determine Role Types of Argument Groups

In this section, I propose heuristics for assigning thematic role labels to the frame-argument configuration groups created as described in the previous section. These heuristics are based on findings concerning linking captured by the linguistic theories which I described in section 2.1. These findings comprise constraints which restrict the range of frame-argument configurations that may realise a certain thematic role type. In addition, the rules proposed here reflect to a large extent the possible syntactic alternations (diathesis alternations) in English. These alternations have been studied in detail in (Levin 1993). Since I have adopted the inventory of thematic roles of EuroWordNet for my work, I concentrate on those linguistic facts which are relevant w.r.t. this inventory and the corresponding definitions of the individual roles (cf. section 2.3.1).

In my work, I will concentrate on four roles: Agent, Patient, Instrument, and Location. Recall that the Theme role, which is central to the thematic role theory developed by Gruber and Jackendoff (but whose adequacy is questioned by Dowty) is not part of the EWN inventory. Thus, I will not address this role type. Furthermore, I will not devise heuristics for several role types occurring in EWN, namely Source (SOURCE_DIRECTION), Goal (TARGET_DIRECTION), Direction, and Result. Source and Goal have been largely discussed in the literature that deals with thematic roles. However, as table 5.1 on page 140 (section 5.4) shows, these roles occur in the gold standard with a rather low number of items (40 and 68, respectively). This low number of test instances decreases the reliability of evaluating the learning algorithm's accuracy of acquiring these roles.[9] However, as I will explain below, heuristics to detect Source and Goal would be analogous to those rules which I use to determine the Location role. This role type is represented in the gold standard by much more items (220). Therefore, I decided to concentrate on this type as the "locative" role with the most training instances. The DIRECTION role type in EuroWordNet has been introduced for verb senses which explicitly involve a direction but where the kind of direction (Source or Goal) is unspecified. (Alonge 1996) notes that some verbs in Italian (e.g. "correre" = "to run") have two different meanings (distinguished by different auxiliary selection) where one involves a directional motion ('run from/to a place') and the other one does not ('run around'). A DIRECTION pointer to a general locative noun concept can be used to distinguish these senses. This is a rather idiosyncratic use of a thematic role relation. Not surprisingly, this role type occurs very infrequently in the gold standard (29 times). The RESULT role is not defined at all in the relevant EWN document (Alonge 1996). Furthermore, it is not obvious whether and how the latter two role types could be related to the linguistic theories which I described in chapter 2. Therefore, I did not take them into account.

In short, the heuristics which I employ to label argument groups with the four above-mentioned role

[9] This shortcoming becomes even more serious if one takes into account that a certain percentage of these test instances are cases of incorporation, i.e. implicit arguments which are not expressed by syntactic complements. As explained in section 5.4, such relations cannot be acquired by learning approaches discussed in this thesis. In chapter 7, I will describe simple heuristics to detect such cases and eliminate them from the gold standard. This further decreases the number of test items.

types are based on the following observations:

1. If a sentence contains an Agent, then the Agent is realised as the subject
2. If a Patient is expressed in a transitive sentence, then in general it is expressed as the object
3. A PP expressing an Instrument is usually headed by "with"
4. Likewise, there are certain prepositions which typically signify a Location, e.g. "in", "at", "on", "above", etc.

All these statements refer to syntactic configurations or individual prepositions as signals for certain role types. This kind of information is available in the verb-frame-argument configurations as encoded in the training data, and thus in the argument groups to be labelled. Other findings concerning the distinction of thematic roles refer to the semantics of the complements which realise a certain role. Such observations are employed in stage 3 of my linking strategy (cf. section 6.5). As noted at the beginning of this chapter, the heuristic rules I will present in this section are not always able to assign one unique role type to an argument group; sometimes multiple role types are assigned. The reason for this is that the kind of information available at this point (syntactic and prepositional) is not sufficient to completely determine the appropriate role. The semantic filters employed is stage 3 have the function to indirectly resolve the remaining "role ambiguity".

Note that terms like "in general", "usually", and "typically" used in the enumeration of observations above indicate that these observations (except from 1.) describe tendencies rather than regularities without exceptions. Indeed, there are a couple of transitive verbs where the subject is a Patient; also there are *with*-PPs that do not express an Instrument or *in*-PPs which do not realise a Location. This is the reason why linking rules depending on these statements are necessarily heuristic in nature. Therefore, applying such rules to determine thematic roles of argument groups brings about a certain amount of noise, i.e. a certain amount of data items represented by the argument groups for which the role assignment is erroneous. The semantic filters help to eliminate such kind of noise as well.

In the following subsections, I turn to each of the four role types that I investigate in detail. I will describe how heuristics for role labelling can be developed from the linguistic findings summarised above and address their limitations. At the end of this section, I will put these heuristics together to a complete sequence of rules which I use in my implementation and discuss some general points regarding this stage of the linking approach.

6.3.1 Agent

All three theories of semantic roles described in section 2.1 state that if an Agent is realised in a sentence, then it is expressed as the subject. Regarding the task of assigning role types to argument groups, this fact can be translated into the following rule:

> **Agent heuristic: (version 1)**
> An argument group that does not contain a subject cannot be labelled as Agent.

This is the only rule presented here that has the status of a fact rather than a default.

Of course, the reverse is not true: Not every group containing a subject can properly be classified as Agent. Recall the example of the groups for "break" in section 6.2 (page 165). All three groups contain at least one frame-argument configuration which represents a subject. However, only one of them, g_2, can appropriately be labelled as Agent, while g_1 corresponds to the Patient and g_3 to the Instrument. As we will see below, I employ a syntactic heuristic to determine whether an argument group represents the Patient role. In addition, I propose a semantic filter to distinguish Agents from Instruments and Locations at a later stage. Therefore, the negative rule above can be transformed into a positive heuristic:

> **Agent heuristic: (version 2)**
> An argument group that contains a subject is labelled as Agent unless it is labelled as Patient due to other heuristics.

6.3.2 Patient

Before explaining the heuristics I employ for labelling argument groups as Patient, I have to make a preliminary remark. The Patient role differs from the other role types we discuss in a crucial respect. As we will see in section 6.5, the other roles can be associated with quite homogeneous coarse-grained selectional preferences which are verb-independent. This means that the noun concepts which usually are involved in role relations of a certain type are subsumed by a few general concepts. For example, the AGENT relation typically is connected with noun concepts subsumed by <life_form>, while the INSTRUMENT relation is typically tied to noun concepts that are hyponyms of <inanimate_object>. As these general semantic preferences of different role types are disjunct, they can be used as semantic filters to distinguish role types. However, the PATIENT relation does not have such homogeneous semantic preferences. A Patient can be a life form as well as an inanimate object. For this reason, it is not possible to employ semantic filters as will be introduced in section 6.5 to distinguish the Patient role from other roles. Hence, within my linking strategy, syntactic heuristics as described in this section are the only way to discover Patients. Fortunately, a rather simple rule can be applied to achieve this. In the following pages, I will motivate this rule and investigate to what extent it is able to cope with several widely discussed linguistic facts. This discussion will be much more comprehensive then the analogous considerations for the other role types. For the recognition of those types, syntactic heuristics only perform a more or less rough preselection of noun instances, whereas the final determination is done by semantic filters. The Patient, in contrast, is determined solely by the heuristic introduced in this subsection. This requires a more thorough investigation whether it can adequately fulfil this task.

The general heuristic

(Jackendoff 1990, p. 129) states that if there is a Patient in an English transitive sentence, then it tends to be realised as the direct object. Essentially, this is also a major consequence of Dowty's theory according to which that argument that meets the most P-Patient entailments is realised as object. This finding is illustrated by our "break" example, where the group that corresponds to the Patient (g_1) is the only one that contains objects. (Dowty 1991, 581) points out that there are a few exceptions, i.e. transitive verbs where the Patient is expressed as the subject. These verbs include "receive", "inherit", "undergo", "suffer (from)", and a couple of others. Dowty emphasises that there is only a small number of such verbs. Furthermore, he argues that for some of these verbs, historical meanings

are proved which imply that the subject really is the Agent of the denoted action. For example, beside the "passive" meaning which is common today, "receive" once also had an "active" sense equivalent to 'take or accept something willingly'; likewise, once a possible meaning of "undergo" was 'submit oneself to'.

Interestingly, this ambiguity is reflected in the gold standard as well. Several wordnets in EWN contain a thematic role relation whose English equivalent is <receive#have> AGENT <recipient#receiver>. This corresponds to an "active" sense of "receive" with a volitional receiver. Strikingly, one wordnet encoding this relation (the Italian wordnet) as well contains the analogous relation which captures the "passive" sense, i.e. <receive#have> PATIENT <recipient#receiver>.[10] A volitional sense of "receive" and "undergo" is also captured by AGENT relations from <experience#receive#undergo> to <recipient#receiver> and <victim> (sic!). Some PATIENT relations from <receive#have> correspond to the passive (nonvolitional) sense (like the one mentioned above), e.g. a relation to <bailee>. Other PATIENT relations correspond to the volitional sense, e.g. from <receive#have> to <prize#award>. (Here, the received entity is the Patient, not the receiver.) These examples illustrate that in the gold standard, even for these exceptional verbs, the "active" sense, which implies the correspondence Agent–subject and Patient–object, is encoded to a certain extent. Thus, a default rule like the following seems justified:

> **Patient heuristic: (version 1)**
> An argument group that contains an object is labelled as Patient.

It turns out that this heuristic (with a minor modification which I will discuss below) is largely sufficient to detect Patients. It might appear surprising that such a simple rule is able to cope with a number of issues concerning this role type discussed in the literature. In particular, three kinds of syntactic configuration concerning the linking problem are relevant in this context:

- intransitive sentences
- verbs subcategorising for two objects
- objects which express a role different from the Patient

In the following, I address these points.

Intransitive sentences

One important point is the treatment of intransitive subjects (i.e. the subjects of intransitive sentences). Examples (6.6 b) and (6.7 b) on page 151 show that the subject of an intransitive sentence may be an Agent (as with the verb "eat") or a Patient (as with the verb "close"). As discussed already, examples (6.6) and (6.7) illustrate that for "eat", the nouns which are selected as the intransitive subject tend to correspond to the nouns selected as the transitive subject, while for "close", the nouns selected as the intransitive subject tend to correspond to the nouns selected as the object. Hence, in an LSC

[10] In a few cases, a verb concept and a noun concept are related via two role relations of different type in the gold standard. Mostly, these different relations originate from different monolingual wordnets. Sometimes, however, one language-specific wordnet provides two role relations of different type which connect the same verb concept and noun concept.

model, *eat#subj/subj* behaves similarly to *eat#subj:obj/subj* w.r.t. the pattern of class membership, while *eat#subj:obj/obj* exhibits a different class membership profile. As a consequence, grouping the arguments of "eat" yields a group that contains the intransitive and the transitive subject types, i.e. a group like

$g_i(eat) = \{$*subj/subj*, *subj:obj/subj*$\}$

Of course, other frame-argument configurations (e.g. subjects belonging to a subcategorisation frame with some PP) could belong to such a group as well. The crucial point is that the object type *subj:obj/obj* is *not* part of that group. For the verb "close", the situation is different: here, the intransitive subject behaves similarly to the object w.r.t. the selection of nouns. This means that in an LSC model, *close#subj/subj* and *close#subj:obj/obj* exhibit similar class membership profiles and thus grouping yields an argument group that contains both types, e.g.

$g_i(close) = \{$*subj/subj*, *subj:obj/obj*$\}$

It is easy to see that the heuristics mentioned so far are sufficient to assign the correct role types to the two argument groups: As $g_i(close)$ contains an object, the Patient heuristic can be applied to that group and correctly classifies it as representing the Patient. In contrast, $g_i(eat)$ does not involve an object so that the Patient heuristic does not apply. However, since this group comprises several subject types, the Agent heuristic is employed classifying it as corresponding to the Agent, which is indeed the case.

This example demonstrates the benefit of grouping frame-argument configurations for determining the corresponding thematic roles. Considered in isolation, a frame-argument configuration generally might express several roles. However, if frame-argument configurations are assembled in role-specific groups, then the set of possible roles corresponding to one configuration is constrained by the possible roles corresponding to the other configurations in the group. More precisely, a role that might be expressed by one frame-argument configuration is excluded if there is another configuration in the same group that cannot realise that role. Here, the type *subj/subj* might correspond to the Agent or the Patient. If this configuration is grouped with an object configuration, then this rules out the Agent role. If, however, the group to which it belongs does not contain an object (but other subject types), then classifying that group by the Agent role is an appropriate default choice. The crucial point is that the combination of frame-argument configurations in a group provides information for determining the appropriate thematic role which is not available from the single frame-argument configurations alone.

Unfortunately, this discussion also reveals the limitations of the grouping strategy. Obviously, the advantage of combining frame-argument configurations can only be effective if there actually are configurations which constrain the range of possible role types. Regarding the role assignment to intransitive subjects, it depends on the examined verb whether such constraining frame-argument configurations are available. More concretely, an intransitive subject is only recognised as Patient if it is grouped with an object type. This requires that the intransitive frame (where the Patient is realised as the subject) has a transitive alternation where the Patient is realised as the object. So far, we have looked at verbs like "break" and "close" where this is the case. However, other verbs do not show this alternation. Some verbs which express a Patient as the intransitive subject do not have a corresponding transitive frame.[11] (Levin 1993) mentions verb classes which do and verb classes which do not

[11]Note that such verbs have to be distinguished from verbs like "receive" or "undergo" mentioned at the beginning of this subsection. Those verbs *are* transitive but (nonetheless) express a Patient by the subject.

exhibit an *object of transitive = subject of intransitive* alternation. For example, "die" is only used intransitively, and the subject expresses a Patient.[12] For such verbs, the frame-argument configuration *subj/subj* cannot be grouped with an object type, since such a type does not exist here. Hence, our heuristics will fail to assign the Patient role to the group containing the intransitive subject configuration. Fortunately, the range of affected verbs seems limited. Levin lists only three (related) verb classes which solely occur with the intransitive variant of the alternation we are discussing, namely *verbs of appearance* ("appear", "arise", "awake", etc.), *verbs of disappearance* ("die", "disappear", "expire", etc.), and *verbs of occurrence* ("ensue", "eventuate", "occur", etc.). Verbs of these classes realise a Patient as an intransitive subject, but not as an object. However, some of them have an additional agentive interpretation, e.g. "appear", "disappear", "or "vanish" (at least if their subject is animate) so that the classification of their intransitive subject as Agent is justified as well.[13]

Double object constructions

The next point we have to consider is the possibility that a verb subcategorises for *two* objects. This is the case for double object constructions as in

(6.17) Harry gave Sam a book.

Here, the question is whether both objects express a Patient or only one of them (and if so, which one). To answer this question, we have to look at the different kinds of verbs which occur with two objects. Levin mentions a number of verb classes with this property. These verb classes can be divided in two groups, according to the alternations which they typically undergo. The first group comprises verbs which usually imply a change of possession (in the broadest sense, including e.g. the exchange of information), e.g. give verbs ("give", "lend", "sell", etc), send verbs ("send", "mail", etc.), or verbs of transfer of a message ("teach", "tell", "write", etc.). These verbs exhibit the *dative alternation*, i.e. the object adjacent to the verb alternates with a *to*-PP:

(6.18)

a. Bill sold a car to Tom.

b. Bill sold Tom a car.

In this alternation, the first object in b. ("Tom") corresponds to the *to*-PP in a. and denotes the recipient of the entity that is denoted by the other object in a. and b. ("a car"). Which of these arguments is a Patient? The linguistic theories described in section 2.1 imply differing views regarding this issue. According to Dowty's approach, both arguments meet a couple of P-Patient entailments. The recipient is stationary relative to another participant of the action and—at least to a certain degree—undergoes a change of state and is causally affected by another participant. The exchanged entity is causally affected, in a sense undergoes a change of state (a change of possessor), and, for some verbs, is an Incremental Theme (e.g. "write a letter"). (Dowty 1991, p. 576) states that two arguments which have "approximately equal numbers of entailed P-Patient properties" may both be realised as direct objects.

[12] The gold standard contains the corresponding relation: <die> PATIENT <person>.

[13] The restriction to animate subjects will be achieved by a semantic filter; cf. section 6.5 below.

From this point of view, both arguments could be characterised as Patients. Jackendoff analyses such sentences differently. Example (2.31) on page 26 contains the conceptual structure for sentence (6.17) above. According to that structure, "Sam" is the Patient (second argument of AFF) as well as the Goal (argument of TO), while "a book" is the Theme (first argument of GO). Hence, the two theories agree in treating the recipient as Patient, but disagree w.r.t. the "Patienthood" of the exchanged entity.

In view of these different analyses, the crucial guide to decide the question is the definition of the PATIENT relation in EWN. In section 2.3.1, I mentioned that the relation "Y INVOLVED_PATIENT X" is defined by the test sentence template "(A/An) X is the one/that who/which undergoes the Y." As noted, the Theme role is not a part of the EWN inventory. However, in the cases discussed here, the definition for the Patient clearly holds for the argument that Jackendoff classifies as Theme, i.e. the entity which is exchanged, since this entity undergoes the exchange action. Furthermore, one can argue that the target of the exchange act (expressed by a direct object or a *to*-PP, respectively) undergoes this act as well, at least to the extent to which the exchange is of relevance for that target. In this regard, an observation noted by Levin is relevant: The alternation illustrated by (6.18) only is possible if the recipient denotes an animate being or an institution, as shown in (6.19):

(6.19)

a. Bill sent a package to Tom/London.
b. Bill sent Tom/*London a package.

It appears that this restriction can be understood in terms of affection of the recipient by the event. If a package is sent to a human (or an organisation), then this is relevant for the target (the recipient). Hence, the target in a sense undergoes the act of being sent a package. However, if a package is sent to a certain location (e.g. a city), then this does not significantly affect that location.[14] (Indeed, the target cannot be referred to as the recipient.) Here, it would be inappropriate to say that the target undergoes the sending action.

Regarding the task of labelling argument groups, it turns out that the Patient heuristic as proposed above adequately covers the facts I have just discussed. We have seen that both objects in the double object variant of the dative alternation are properly classified as Patient. This is exactly what the Patient heuristic does, since it assigns the Patient role to any object. In particular, it does not distinguish between the two objects in a double object frame. Moreover, it treats the *to*-PP in an appropriate manner. This PP expresses the target of the exchange action. As shown above, this target can only be viewed as a Patient (i.e. as something which undergoes the delivery action) if it also can be expressed as object, i.e. if the dative alternation is possible. In this case, the respective frame-argument configurations *subj:obj:obj/obj2* and *subj:obj:pp.to/pp.to* should be member of the same argument group, which the Patient heuristic labels as Patient. If, however, the dative alternation is not possible, then the group that contains *subj:obj:pp.to/pp.to* should not contain *subj:obj:obj/obj2* (nor any other object type) and thus is not classified as Patient.

The other alternation mentioned by Levin that involves a double object construction is the *benefactive alternation*. In this alternation, the first object of a double object construction alternates with a *for*-PP, and this alternating argument expresses a Beneficiary of an action, as in

[14] Of course, this does not apply for a metonymic interpretation where the name of a location denotes a group of people or an organisation, e.g. if "London" refers to a firm's office in London. In case of such an interpretation, the dative alternation is possible.

(6.20)

a. Martha carved a toy for the baby.

b. Martha carved the baby a toy.

This alternation occurs, broadly speaking, with verbs of creation (e.g. “build”, “design”, “bake”, or “write”) or get verbs (e.g. “buy”, “catch”, or “pick”). Again, the question here is whether one or both objects realise a Patient. Following the EWN definition for the PATIENT relation, I think it is uncontroversial that the non-alternating object (“a toy” in this example) expresses a Patient. This object expresses the entity which is created or obtained, respectively. One can say that a toy undergoes the action of being carved as well as a flower undergoes the action of being picked. However, in my opinion, the alternating object requires different treatment than the alternating object in the dative alternation. It is rather odd to say that the baby undergoes the action of being carved a toy or the girlfriend undergoes the action of being picked a flower. Thus, an object that expresses a Beneficiary of a verb exhibiting the benefactive alternation should not be classified as Patient. One can account for that by extending the Patient heuristic by a simple exception: If an argument group containing an object configuration also contains a *for*-PP configuration, then this group must not be labelled as Patient. This exception is based on the reasonable assumption that for verbs which undergo the benefactive alternation, the types *subj:obj:obj/obj2* and *subj:obj:pp.to/pp.for* are grouped together. Therefore, we have to revise the heuristic accordingly:

Patient heuristic: (version 2)
An argument group that contains an object is labelled as Patient unless it contains a *for*-PP.

One might object that the difference in Patienthood of the recipient in the dative alternation and the beneficiary in the benefactive alternation is rather degreed than categorical. Indeed, my decision to treat the former as a Patient and the latter not is debatable. However, the general linking approach I propose is open to other decisions in this respect. If the dative recipient should not be classified as Patient either, then the heuristic has to be modified in the way that the Patient assignment is also precluded if the argument group contains a *to*-PP. If, on the other hand, the beneficiary of benefactive verbs should be treated as Patient as well, then version 1 of the Patient heuristic has to be applied without exceptions.

Locative alternations

Apart from double object constructions, there is a further possibility that two arguments of a verb are realised as object. Levin discusses several alternations which comprise two subcategorisation patterns each of which includes an object, but the objects in the different patterns contain different kinds of participants of the event denoted by the verb. The best-known alternation with these characteristics is the *spray/load alternation*, which I have mentioned several times in sections 2.1 and 6.2.3. Many verbs expressing some kind of putting or covering undergo this alternation. I will discuss it as representative of similar cases which Levin subsumes as *locative alternations*. The arguments I provide apply analogously to other alternations of that sort.

Consider the following example:

(6.21)

a. Bill loaded the books onto the truck.

b. Bill loaded the truck with the books.

In terms of Jackendoff's thematic tier, the object in sentences like (6.21 a.) expresses the Theme (the entity which is moved), while the object in sentences like (6.21 b.) expresses the Goal of the described action. Jackendoff and Dowty state that in both cases it is the object that expresses the Patient. Dowty justifies that claim by a subtle argument based on proto entailments (I sketched that argument in section 2.1.3). Jackendoff provides a simple test for the Patient role. If the sentence in question can be paraphrased using the template *"What happened / What Y did to NP was..."*, then *NP* represents a Patient. Applying this test to the sentences in (6.21) yields the following:

(6.22)

a. (i) What Bill did to the books was load them onto the truck.
   (ii) ?What Bill did to the truck was load the books onto it.

b. (i) What Bill did to the truck was load it with books.
   (ii) *What Bill did to the books was load the truck with them.

Thus, "the books" is the Patient in (6.21 a.) and "the truck" in (6.21 b.). The essential observation underlying these arguments is that the sentences in (6.21) have a slight difference in meaning. Roughly speaking, this difference consists in which of the non-agentive participants is "more directly affected" by the action. In (6.21 a.), it is "the books", while in (6.21 b.), it is "the truck". Such subtle meaning differences are usually not reflected by sense distinctions in (Euro)WordNet.[15] Moreover, a thematic role relation in EWN is a concept-to-concept relation. It is not possible (and not intended) to specify that a role relation is related to a specific subcategorisation frame of the verb. Therefore, the information that "the books" (or "the truck", respectively) can only be the Patient of "load" if expressed as object cannot be encoded in EuroWordNet.

The EWN definition for PATIENT fits for both arguments in question; it is adequate to say that the books as well as the truck undergo the loading act. Thus, for a verb which exhibits the spray/load alternation, a PATIENT relation is appropriate to concepts which denote either the "Theme" or the "Goal" of the action expressed by that verb. Here, again, the Patient heuristic works appropriately. Due to the alternation, there should be several argument groups which contain an object type: the "Theme" is represented by a group containing an object and a *with*-PP, whereas the "Goal" is represented by a group containing an object and a locative PP. In section 6.2.3, we have seen that these groups really are acquired. The heuristic classifies both as Patient, which is the desired behaviour.

Levin also mentions verbs related to putting or covering which do not participate in the spray/load alternation. Some of such verbs only occur with the variant involving a *with*-PP:

(6.23)

a. *Marc filled water into the bowl.

[15] If such distinctions were encoded, then, for example, the respective meaning of "load" in (6.21 a.) and (6.21 b.) should be represented by two different synsets. Actually, this is not the case in WordNet.

b. Marc filled the bowl with water.

Other verbs only appear with the variant containing a locative PP:

(6.24)

a. Tamara poured water into the bowl.
b. *Tamara poured the bowl with water.

For the former kind of verbs, my linking strategy only classifies the "Goal" (here, "the bowl") as Patient; for the latter kind of verbs, only the "Theme" (here, "water"). One could argue that, according to the PATIENT criterion of EWN, the other non-subject arguments are Patients as well: The water undergoes the filling, the bowl undergoes the pouring. The Patient heuristic which I propose does not recognise these arguments, because they are never realised as object. In this sense, this heuristic misses some arguments that express a Patient. To capture these arguments, the heuristic would have to be extended in a way that it also labels an argument group as Patient if it contains a *with*-PP or a locative PP. However, this would raise serious problems regarding the distinction between Patients and other role types. A *with*-PP is a strong indicator of the Instrument role; likewise, locative PPs indicate a locative role type. Thus, groups comprising these PPs would have to be classified as *Patient or Instrument* (in case of a *with*-PP) or as *Patient or locative role*[16] (in case of a locative PP), and semantic filters would be necessary to decide between these alternatives. However, as mentioned at the beginning of this subsection and as I will show in more detail in section 6.5, the semantic range of the nouns occurring as Patient overlaps with the semantic ranges of the other role types so that it is not possible to distinguish the Patient by semantic filters. Therefore, to obtain unique role assignments at the end of the linking process, I decided to restrict the Patient heuristic to objects.

This restriction is justifiable by linguistic reasons. (Fillmore 1977) addresses the issue of realising arguments as object or PP, respectively, in terms of saliency (cf. section 2.1.1). His theory states that the arguments which are most salient in the scene described by a sentence are expressed as subject and object, while other arguments are realised as PPs (or other constituents). Different verbs put different participants of an event into the foreground. For example, "fill" emphasises the container (it implies that the container is full after the event), while "pour" stresses the movement of the liquid. Dowty and Jackendoff provide analogous (though more sophisticated) considerations within their theories. Roughly speaking, these considerations justify to solely regard the respective (possible) objects of the individual verbs as Patients (or, at least, ascribe them a "higher Patienthood") than the PPs. Thus, concentrating on the objects is reasonable, since in this way, the "salient" or "primary" Patients can be discovered.

Interference with other roles

Finally, another issue has to be addressed: Is it possible that other roles are expressed as object? If this is the case, then the heuristic I propose would wrongly classify such cases as Patient. As we have seen in section 6.3.1, the Agent does not "compete" with the Patent in this respect, since this role is always realised as subject. Regarding the remaining roles, I will turn to that question in the respective sections 6.3.3 and 6.3.4.

[16] I will address the issue of distinguishing locative roles (Location, Source, or Goal) in section 6.3.4.

6.3.3 Instrument

During the current discussion, we have seen various examples in which the Instrument is realised either as the subject or as a *with*-PP, as in

(6.25)

a. David broke the window with the hammer.
b. The hammer broke the window.

Levin refers to this pattern as the *Instrument subject alternation*. She notes that not all kinds of Instruments can be expressed as subject. More precisely, there is a distinction between *intermediary* and *enabling/facilitating* instruments. Only the former kind can appear in the subject position:

(6.26)

a. The crane loaded the truck. (intermediary Instrument)
b. *The pitchfork loaded the truck. (enabling/facilitating Instrument)

Moreover, there are verbs that only select enabling/facilitating Instruments, e.g. “eat” or “see”:

(6.27)

a. Doug ate the ice cream with a spoon.
b. *The spoon ate the ice cream. (enabling/facilitating Instrument)

All these cases have in common that a *with*-PP signals an Instrument. This motivates a preliminary formulation of a heuristic for labelling argument groups that express an Instrument:

Instrument heuristic: (version 1)
An argument group that contains a *with*-PP is labelled as Instrument.

In a sense, this is the core of the Instrument heuristic. However, this rule requires an extension, which concerns the subject. As the examples under (6.26) illustrate, a verb might exhibit a difference between those Instruments which are realised as a subject and those Instruments which are expressed by a *with*-PP. If this difference is (quantitatively) significant, then it might happen that the *with*-PP type and the corresponding subject type are not members of the same argument group. If this is the case, however, then the heuristic stated above will not capture those Instrument instances which are expressed as subject. To overcome that shortcoming, I revise the Instrument heuristic as follows:

Instrument heuristic: (version 2)
An argument group that contains a *with*-PP or a subject is labelled as Instrument.

In this way, Instruments at the subject position are captured even if the corresponding frame-argument configuration is not in a group that also contains a *with*-PP. An obvious objection to this extension is that it is too permissive. If *any* subject type is classified as Instrument, then those subjects that realise a different role are misclassified. In particular, since Agents are always expressed as subject, all argument groups that represent an Agent misleadingly receive the Instrument label. Actually, the problem is that the syntactic information encoded in argument groups is not sufficient to distinguish Agents and Instruments. However, as noted already, semantic criteria are able to perform this distinction. Therefore, the solution of this problem is to assign both role types Agent and Instrument to groups that contain a subject and to leave the final discrimination to the semantic filters.

It is important to note that the presence of the *with*-PP type in an argument group does not exclude the possibility that this group represents an Agent. I emphasise this point because there is a manifest consideration implying the contrary: As the Agent is always realised as the subject, an argument group which contains a subject and a *with*-PP cannot correspond to the Agent role. Therefore, it would be adequate not to assign the Agent, but only the Instrument label to such a group. Unfortunately, that would yield inappropriate results. Although a *with*-PP never realises an Agent immediately, it is quite common that it expresses an argument which is of the same kind as the Agent. This phenomenon is closely related to an alternation which Levin calls the *simple reciprocal alternation* for intransitive verbs. Consider the following example:

(6.28)

a. Brenda cooperated with Molly.
b. Brenda and Molly cooperated.

Here, (6.28 a.) contains a subject and a PP, whereas in (6.28 b.), both arguments are coordinated and appear in the subject position. In both variants, the subject expresses the Agent role. This alternation makes overt the fact that for certain verb classes (e.g. correspond verbs, meet verbs, talk verbs, or amalgamate verbs), a PP which usually is headed by "with"[17] expresses an argument that also could occur at the subject position, where in many cases it corresponds to an Agent. For such verbs, it is likely to obtain an argument group that contains the intransitive subject as well as the *with*-PP type. It is inadequate to classify these groups solely as Instrument and exclude the assignment of the Agent role, which often would be the correct role type. Therefore, my strategy invariably classifies an argument group containing a subject and a *with*-PP as both Agent and Instrument and leaves the final decision on the correct role to the semantic filters.

Semantic filters are also necessary to eliminate noise. In many cases, a *with*-PP neither expresses an Instrument nor an argument that is coequal to the Agent, e.g. in

(6.29) The vase broke with a loud bang.

Such cases can only be eliminated by semantic filters.

To complete the discussion of the Instrument heuristic, I will shortly consider the possibility that other argument slots apart from the subject and the *with*-PP could express an Instrument. Two kinds

[17] A few verbs exhibit the described behaviour involving other prepositions, e.g. "to" ("join") or "from" ("divorce").

of arguments have to be considered: objects and PPs headed by a preposition different from "with". Concerning the former, it is a linguistic finding that an Instrument is normally not realised by an object. (Jackendoff 1990, p. 259) explicitly excludes the Instrument role in that part of his linking theory which is concerned with NP complements, i.e. subjects and objects. (He analyses instrumental subjects as in (6.25 b.) as inanimate Agents.) He mentions one exceptional case of a verb whose object expresses an Instrument, namely "use". Nevertheless, Levin lists a number of verbs where the object and a *with*-PP alternate, as in

(6.30)

a. I mixed the sugar with the butter.

b. I mixed the sugar and the butter.

This is the transitive variant of the reciprocal alternation illustrated above by example (6.28). Here, the two arguments expressed by the object and the *with*-PP as in (6.30 a.) can be coordinated and both realised as the object, as in (6.30 b.). Furthermore, both arguments express Patients, not Instruments. Regarding the grouping of frame-argument configurations, this alternation is expected to result in a group containing the object and the *with*-PP type. This group would be labelled as Patient according to the Patient heuristic and as Instrument according to the Instrument heuristic. However, in this case only the Patient classification is correct, since an Instrument generally is not expressed by an object. Therefore, the Instrument heuristic has to be modified to prevent classifying a group as Instrument if it contains an object:

Instrument heuristic: (version 3)
An argument group that contains a *with*-PP or a subject is labelled as Instrument, unless it contains an object.

Finally, I have to address the issue whether an Instrument can be expressed by a PP not headed by "with". In fact, there are some alternative prepositions which come into question. (Gruber 1965, p. 139) notes that a PP headed by "by" or "by means of" can express an abstract Instrument, as in

(6.31) Bill turned John into a pumpkin by magic.

Unfortunately, "by" is ambiguous. Apart from an instrumental reading, a *by*-PP in an active sentence can express other things, e.g. local proximity ("stand by me") or the quantity of a degree ("increase by 5 per cent"). To decide whether it is suitable to employ this preposition as Instrument marker, I manually examined the argument groups containing a *by*-PP which I acquired from the training corpus (cf. section 6.2.3). The result was that such groups existed for 16 verbs. For 4 verbs, these groups represented an (abstract) Instrument. This means that firstly, a *by*-PP complement is rare, and secondly, it is a bad predictor for the Instrument role. I obtained similar findings for the preposition "through", which might also signal an abstract instrument, but is used to indicate a path in most cases. Therefore, I did not take prepositions other than "with" into account to detect Instruments.

6.3.4 Location

This subsection addresses the Location role. This role is closely related to the widely discussed role types Source and Goal. All three types are concerned with locative categories; therefore, I will refer to them as *locative role types*. Their commonality as well as their differences can be easily revealed by comparing the definitions in EWN mentioned in section 2.3.1. A LOCATION relation from verb Y to noun X in EWN means: "(A/An) X is the place *where* Ying happens / one Ys." The analogous definition of the SOURCE_DIRECTION pointer is: "(A/An) X is the place *from which* Ying happens / one Ys.", whereas the TARGET_DIRECTION (corresponding to the Goal) is characterised by: "(A/An) X is the place *to which* Ying happens / one Ys." In other words, Source and Goal specify a place where the action specified by the verb starts or where it ends, respectively. Furthermore, these definitions presuppose that not the complete action takes place at that location. This implies that this action involves a kind of motion. Note that the term "the place" in these definitions restricts the two role types to their primary locative domain. This means that verbs describing an abstract motion such as a change of possession (which does not necessarily imply a motion in space) are not supposed to select a Source or a Goal. This differs from analyses in (Gruber 1965) and (Jackendoff 1990) (cf. section 2.1.2). Following them, e.g. "sell" has a Source (the seller) and a Goal (the buyer). In general, these participants of the denoted action are captured in EWN by the Agent and Patient role, respectively.

In contrast to Source and Goal, the Location role indicates a place where typically the complete action denoted by the verb happens. This role type is not restricted to verbs denoting motion. Stative verbs involve a Location as well. For motion verbs, the Location delimits the space within which the denoted action takes place. Beside these differences, Location, Source, and Goal specify a location where an action (partly or completely) happens. Hence, it is not surprising that the linguistic means to express them are similar and overlapping to a large extent. Therefore, the following paragraphs, while focusing on Location, will also include aspects concerning Source and Goal. However, I will not discuss the latter two role types systematically. The main reason for this is that they are not included in the evaluation described in chapter 7. At the beginning of section 6.3, I have justified this exclusion by practical shortcomings of the gold standard. This subsection will show that Source and Goal would have to be treated in an analogous way as the Location role. On the other hand, we will see that it is very hard to distinguish them, in particular Location and Goal.

Locative role types are usually expressed by PPs which are headed by certain prepositions. Unlike for the Instrument role, there are multiple prepositions that indicate a locative role. In particular, prepositions used to express a Location include "in", "at", "on", "under", "above", or "around". In contrast, typical prepositions that signal a Goal are "to", "into", or "onto". A Source is usually indicated by "from", but also by "out of" or "off of". Typically, these roles are exemplified in the literature by sentences like

(6.32) Mare travelled from Augsburg to Cologne in a comfortable train.

where "from Augsburg" denotes the Source, "to Cologne" the Goal, and "in a comfortable train" the Location.

This suggests the following heuristic to detect the Location role:

Location heuristic: (version 1)
An argument group that contains a PP headed by a locational preposition (“in”, “at”, “on”, “under”, “above”, “around”,...) is labelled as Location.

In principle, analogous heuristics (including the appropriate prepositions) can be formulated for Source and Goal.

Not surprisingly, as for other role types, certain linguistic facts complicate the situation. One obvious fact is that the PPs mentioned might represent other arguments than a Location:

(6.33) The package will be delivered in two weeks.

(6.34) This rule does not work in many cases.

(6.35) We have to address this issue in a certain way.

(6.36) The workshop participants had lunch at 2 p.m.

(6.37) He went on a long journey.

(6.38) The baby weighs over seven pounds.

(6.39) Lots of assistants work under horrible conditions.

Note that this also holds for prepositions used to express a Source or a Goal:

(6.40) Several hundred people died from SARS.

(6.41) The vase broke into a thousand pieces.

It would take us too far to discuss the nature of non-locative entities which may co-occur with a locative preposition.[18] The crucial point is that, again, semantic filters can be employed to recognise locative PPs which really express some kind of locative entity. I will describe the filters that I use for this in section 6.5.

However, there is a problem which cannot be solved by semantic filters: the distinction between Location and Goal can be difficult. Actually, this problem does not arise for my practical work, since I do not take Source and Goal into account. Nevertheless, since this decision has been made for purely pragmatic reasons concerning the available resources, I mention that issue here. Gruber and Jackendoff point out that for motion verbs, locative PPs may be ambiguous regarding whether they express a Goal or a Location. For example, in

(6.42) The ball rolled behind the house.

[18] For example, (Lakoff & Johnson 1980) offer a theory that explains such observations in terms of metaphor.

"behind the house" may express the target of the rolling event or delimit the place where the complete event happens. This ambiguity occurs with many locative prepositions, such as "behind", "in front of", "above", or "under". Some prepositions are not ambiguous in this way: "in" and "on" indicate a Location, whereas their counterparts "into" and "onto" signal a Goal. Thus, one could formulate heuristics that label argument groups containing one of the former two prepositions as Location and argument groups containing one of the latter two as Goal. However, if an argument group only contains ambiguous locative prepositions, it is not obvious which role should be assigned. Several strategies are imaginable and justifiable to cope with these cases. One could assign such groups the Location role, because the Goal only occurs with motion verbs, while the Location is not restricted to that verb class. Or one could label them as Goal, because a Location does not belong to the argument structure of most verbs, but rather has the function of a restrictive modifier of the denoted event (cf. the analyses in (Jackendoff 1990, p. 72)). Or one could treat motion verbs and other verbs differently by choosing Goal for the former and Location for the latter. This, however, would raise the problem of automatically recognising motion verbs. Obviously, semantic filters are not helpful to distinguish the two role types, since both of them select locative entities, i.e. entities of the same semantic sort. Anyway, a heuristic solution of that problem is possible. However, since I do not take the Goal into account for independent reasons, I do not have to decide which alternative to pursue.

There is a further subtle differentiation which is made in Jackendoff's theory of semantic structures. He distinguishes the major semantic categories Place and Path. These categories are illustrated in the following examples:

(6.43)

a. The plane flew around over the city.
b. The plane flew over the city towards the mountains.

In (6.43 a.), "over the city" delimits the place where the flying event happened. (6.43 b.) implies a directional flying motion, and "over the city" denotes the path of that motion. Again, prepositions like "over" are ambiguous w.r.t. that differentiation. I decided to regard both cases as expressing a Location, because in my view, both meet the EWN definition of Location (although the expression of a place is more "prototypical" in this respect than the expression of a path). This decision affects the set of prepositions to be included in the Location heuristic. For example, "through" expresses a path, but not a place. As we will see below, I included this preposition in the final version of the heuristic.

As for the Instrument role, I have to address the question whether a Location can be also expressed by the subject or the object of a verb, rather than a PP complement. Concerning the subject, some of the alternations presented by Levin involve a correspondence between a locative PP of one syntactic variant and the subject of the other variant. One of these is the *swarm alternation*, illustrated by (6.44):

(6.44)

a. Bees are swarming in the garden.
b. The garden is swarming with bees.

This alternation holds for a number of verb classes, e.g. for verbs of emission of light, sound, or substance, or for swarm verbs. Another alternation, exhibited by a rather small verb class called fit verbs, is shown in (6.45):

(6.45)

a. We sleep five people in each room.

b. Each room sleeps five people.

If such an alternation is reflected by the data pertaining to a particular verb, then subject and locative PP will be in the same argument group and the Location heuristic will label this group as Location. However, such a group is also labelled as Agent due to the Agent heuristic. This is analogous to ambiguous groups w.r.t. the distinction between Agent and Instrument. As for that distinction, semantic filters are able to distinguish between Agent and Location. However, the situation is different for the distinction of Instrument and Location. We will see in section 6.5 that it is virtually not feasible to discriminate these two roles by semantic filters. The reason for this is that many entities can be referred to as Instrument or as Location. For example, a chair can be viewed as an instrumentality (e.g. for decorating rooms or, sometimes, for breaking windows) or as a Location where a person sits or a thing rests on. Therefore, we have to formulate the heuristics in a way that an argument group is not assigned both the Instrument and the Location type. To achieve this, I revise the Instrument heuristic in the following way:

Instrument heuristic: (version 4)
An argument group that contains a *with*-PP is labelled as Instrument, unless it contains an object. In addition, an argument group that contains a subject is labelled as Instrument, unless it contains an object or a locational PP.

This refinement has the effect that the classification of an argument group as Instrument due to a subject is blocked if this group also contains a locative PP (which triggers the assignment of the Location type). Note that the above formulation of the rule does not impose that restriction on a group containing a *with*-PP. Levin does not mention any alternation between a *with*-PP and a locative PP. Therefore, these PPs should not occur in the same argument group so that a conflict should not arise. In case a *with*-PP and a locative PP do co-occur in a group, there is an erroneous grouping. In such a case, I allow a simultaneous labelling of this group as Instrument and Location, because there is no obvious way to decide which of these roles would be more appropriate.

There is also an alternation between a locative PP and the object. This alternation, called the *preposition drop alternation*, involves a PP expressing a path (cf. (6.46)) or a Source (cf. (6.47)):

(6.46)

a. They skated along the canals.

b. They skated the canals.

(6.47)

a. They escaped from the prison.

b. They escaped the prison.

Again, if this alternation occurs significantly for the verb under examination, then argument groups are built which contain the object and the locative PP. The Location heuristic in the above-mentioned version (or an analogous Source heuristic, respectively) would label such a group as Location (Source). And again, as above, this role assignment would "compete" with a different role assignment due to another heuristic, in this case the Patient heuristic. But unfortunately, semantic filters are not capable of distinguishing Patients and locative roles, which is the consequence of the broad semantic variety of nouns that occur as Patients. In fact, locations can be Patients, as in

(6.48) The caretaker swept the street.

(6.49) Mary decorated the room.

In these examples, the objects denote locations, but the thematic role they express in first place is the Patient. Of course, they can also be regarded as expressing both Patient and Location. In Jackendoff's theory, the former role is part of the action tier, while the latter one rather corresponds to the thematic tier. However, since the semantic representation of roles in EWN does not include the distinction of different tiers, my linking approach assumes that each argument expresses only one role. In (6.48) and (6.49), the Patienthood of the objects is salient. In other, syntactically analogous, cases, the situation is different. Several motion verbs incorporate locative prepositions so that locative roles obligatorily appear as an object rather than a PP. This means that, unlike in (6.46) and (6.47), the objects of such verbs do *not* alternate with locative PPs. For example, the object in (6.50) expresses a Location (a path), in (6.51) a Goal, and in (6.52) a Source:

(6.50) The pedestrians traversed the street.

(6.51) John entered the room.

(6.52) George left the room.

In these sentences, the Patienthood of the objects, if present at all, is secondary. (In my view, it is rather odd to say that e.g. the room undergoes the entering action.) In other words, for these verbs, the Patient heuristic inappropriately assigns objects the Patient role. Examples (6.48)–(6.52) show that it depends on the verb whether its object has to be classified as Patient or as a locative role.[19] To automatically distinguish cases like (6.48) and (6.49) from cases like (6.50)–(6.52), information about the semantics of the individual verbs is required which is not overtly available in WordNet. Therefore, I decided that the rules in my approach do not depend on the particular verb processed. Hence, the linking rules have to treat both kinds of verbs in the same way; a heuristic like "Label an argument group which contains an object as Patient if the verb is like ... and as Location if the verb is like ..." is not possible. On the other hand, the preliminary assignment of multiple roles, i.e. the Patient and locative roles, is not possible either, since semantic filters cannot distinguish between them. Hence, the linking heuristics have to classify argument groups consisting of objects either as Patient or as Location. Obviously, the first alternative is the adequate choice, since objects express Patients for the vast majority of verbs.

But how should verbs be treated which *do* exhibit an alternation between an object and a locative PP, as in (6.46) and (6.47)? Does the locative PP provide sufficient evidence to resolve the Patient/locative

[19] Furthermore, examples (6.50)–(6.52) show that in the latter case, the question *which* locative role is realised by an object is also verb-dependent.

ambiguity, i.e. to exclude the possibility that the corresponding argument group represents a Patient? If this were the case, then the Patient heuristic should be modified in a way that an argument group containing an object is not classified as Patient if the group also contains a locative PP. Unfortunately, this would not be appropriate for all verbs either. In particular, verbs which undergo the *conative alternation* (e.g. hit verbs, cut verbs, some spray/load verbs, or push/pull verbs) express a Patient either by an object or by a PP headed by a locative preposition (typically "at"):

(6.53)

a. I pushed the table.
b. I pushed at/on the table.

Note that there is a meaning difference between (6.53 a.) and (6.53 b.). The first variant implies a causal affection of the second argument, while the second variant does not. (Here, (6.53 a.) implies that the table has been moved; (6.53 b.) does not imply that.) Consequently, Dowty and (Fillmore 1977) discuss whether and to what degree the second argument has to be regarded a Patient at all. Anyway, both variants pass the EWN test for Patienthood: The table undergoes the pushing action. Hence, in these cases, it is appropriate to classify the corresponding argument groups as Patient. This means that the occurrence of a locative PP in a group also containing an object is no reliable indicator for the corresponding locative role.

As a conclusion of this discussion, I decided to generally give preference to the Patient in case of a Patient/Location ambiguity. This can be achieved by modifying the Location heuristic so that it does not label an argument group as Location if it contains an object. Such a group will solely be labelled as Patient by the Patient heuristic. Thus, the final version of the Location heuristic is as follows:

Location heuristic: (version 2)
An argument group that contains a PP headed by a locational preposition, i.e. "in", "inside", "at", "on", "above", "under", "below", "behind", "before", "around", or "along", is labelled as Location unless it contains an object.

Note that this version also lists all prepositions which I take into account as indicators of Location.

6.3.5 Summary and Concluding Remarks

At the end of this section, I list the final versions of the heuristics for labelling argument groups with the four roles I take into account, i.e. Agent, Patient, Instrument, and Location:

Agent heuristic:
An argument group that contains a subject is labelled as Agent, unless it contains an object.

Patient heuristic:
An argument group that contains an object is labelled as Patient, unless it contains a *for*-PP.

Instrument heuristic:
An argument group that contains a *with*-PP is labelled as Instrument, unless it contains an object. In addition, an argument group that contains a subject is labelled as Instrument, unless it contains an object or a PP indicating a Location (as listed in the Location heuristic).

Location heuristic:
An argument group that contains a PP headed by a locational preposition, i.e. "in", "inside", "at", "on", "above", "under", "below", "behind", "before", "around", or "along", is labelled as Location, unless it contains an object.

The discussion of these heuristics showed that they adequately deal with a number of syntactic alternations which involve realisations of a particular role by different syntactic arguments within different subcategorisation frames. I also pointed out cases where subtle differentiations which are not captured by the heuristics would be necessary for an appropriate role assignment. Such differentiations are not possible within the framework which I set up for my linking strategy at the beginning of this chapter, which, in turn, is constrained by the general setting of this work (the ultimate task, available resources, etc.). For example, I abandoned the possibility to formulate heuristics dependent on the kind of verb that is actually processed. We have seen that under certain circumstances, this would be required to determine the correct role (e.g. to discriminate between Patient and Location). However, such differentiations would require information about the semantics of individual verbs that is not available. One could imagine to develop methods to obtain such verb-specific information from the information present in the wordnet (e.g. the glosses of verbs or their location within the hierarchy). Such strategies would highly depend on the specific wordnet under consideration. However, this thesis aims at developing approaches which are as language-independent as possible. Thus, the intended portability of the proposed linking approach is accompanied by limitations in accuracy, which is not surprising. Nevertheless, it seems remarkable to me that one rather simple heuristic for each role type is able to cover a considerable variety of alternations related to the realisation of the respective role.

In section 6.1, I pointed out that the role assignment heuristics described here constitute the part of my linking approach that is predominantly language-dependent. The 1. stage of this approach, building syntactic argument groups, is its most universal part; it neither depends on the employed inventory and definition of roles, nor on the language for which role relations are to be acquired. The semantic filters applied in the 3. stage for the final role discrimination are primarily determined by the assumed role inventory, particularly on the definition of the semantic characteristics of the individual roles. The heuristics discussed in this section, which are applied in the 2. stage, also depend on this factor. When devising the collection of heuristic rules, I anticipated the capability of semantic filters to discriminate different role types: The assignment of multiple roles to an argument group is only admissible if these concurrent roles can be distinguished by semantic filters. However, language-dependency is a particular property of the role assignment heuristics. Unlike the other stages, these heuristics are to a large extent specific to English. If another language had to be processed, then the rules would have to be adapted or even substantially revised.

Let me briefly mention some exemplary points subject to necessary modifications of the heuristics. First of all, it is obvious that at least the prepositions specified in the rules would have to be replaced by corresponding prepositions, postpositions, case markings etc. which express the respective role in a different language. For example, an Instrument is expressed by a PP headed by "mit" in German and "avec" in French; in German, the case of the NP inside a locative PP determines whether this PP

expresses a Location or a Goal (accusative indicates a Goal as in "auf die Straße", dative a Location as in "auf der Straße"). This example also illustrates that the discriminative accuracy of role assignment heuristics strongly relies on the means used by the examined language to express semantic roles. Since in German the Goal/Location ambiguity discussed in section 6.3.4 (cf. example (6.42) on page 186) is resolved by the case inside the PP, these roles can be easily distinguished by heuristics. Another question which has a strong impact on the formulation of role assignment heuristics is which syntactic alternations occur in the examined language. For example, in French, the dative alternation, which typically is exhibited by change of possession verbs in English (cf. example (6.18) on page 177), is only possible if the recipient is expressed by a pronoun. Otherwise, the recipient is realised by a PP rather than an object ("Jean donne le journal *a ses amis*"). If, as I have argued, the recipient is regarded as a Patient, then the restriction of the Patient heuristic to objects is not sufficient; it has to be extended to also capture argument groups containing PPs corresponding to recipients.

Finally, as Dowty and (Fillmore 1968) noted, ergative languages require a substantial modification of role assignment rules. Ergative languages (e.g. Tibetan) exhibit a relation between cases and underlying roles which in a sense is reverse to non-ergative languages like English. In English, a transitive sentence usually expresses the Agent as subject and the Patient as object. An intransitive sentence only has a subject, which may express an Agent or a Patient. In ergative languages, a transitive sentence expresses the Agent by an argument bearing the ergative case and the Patient by an argument bearing the absolutive case. An intransitive sentence realises an Agent or a Patient by an absolutive complement, i.e. by an argument which, roughly speaking, corresponds to an object in non-ergative languages. To appropriately deal with this constellation, the Agent and the Patient heuristic would have to be exchanged: An argument group that contains an ergative has to be labelled as Agent, while an argument group that contains an absolutive has to be classified as Patient, unless it contains an ergative.

The dependency of role assignment heuristics of the kind presented here to a specific language reflects the language-dependency of linguistic linking theories in general. In other words, due to considerable variations across individual languages and language families, it is not possible to develop a linking strategy which is completely language-independent. Unfortunately, this fact is opposed to the desideratum that the solutions proposed in this thesis should not be bound to a particular language. However, regarding this desideratum, it is an advantage of my linking strategy that language-specific aspects are isolated within one module. This facilitates the adaption to other languages. (Jackendoff 1990, p. 261f) poses the question which aspects of a linking theory are universal and which parameters guide language-dependent variations. To the extent to which linguistic research will be able to answer this question, language-independent and language-specific portions of the heuristics presented here can be made explicit.

6.4 Preparing the Input of the Learning Algorithm

The set of argument groups for a verb which are labelled with role types provides the information required to apply the algorithm for selectional preference acquisition to semantic roles rather than syntactic arguments. This information has to be transformed appropriately to be suitable as input of this algorithm. This section describes that transformation.

As the first step, the argument groups are broken down again into individual frame-argument configurations. However, each frame-argument configuration inherits the role label(s) of the group(s) it is a

argument group	role type(s)
$g_1(break) = \{$*subj/subj*, *subj:obj/obj*, *subj:obj:pp.with/obj*$\}$	Patient
$g_2(break) = \{$*subj:obj/subj*, *subj:obj:pp.with/subj*$\}$	Agent, Instrument
$g_3(break) = \{$*subj:obj/subj*, *subj:obj:pp.with/pp.with*$\}$	Agent, Instrument

Table 6.1: Argument groups for the verb "break" and the corresponding role types

frame-argument configuration	role type(s)
subj/subj	Patient
subj:obj/obj	Patient
subj:obj:pp.with/obj	Patient
subj:obj/subj	Agent, Instrument
subj:obj:pp.with/subj	Agent, Instrument
subj:obj:pp.with/pp.with	Agent, Instrument

Table 6.2: frame-argument configuration – role type mapping for the verb "break"

member of. In this way, individual frame-argument configurations are related to one or more corresponding role types. As an example, let us look at the argument groups of the verb "break" as listed at the end of section 6.2.1. Table 6.1 lists these groups and the roles assigned to them by the heuristics developed in the previous section.

g_1 is labelled as Patient, because it contains an object. g_2 and g_3 are labelled as Agent, because they contain a subject. g_1 also contains a subject, but the Agent heuristic does not apply here because this group also contains an object. Likewise, g_2 and g_3 are labelled as Instrument, because they contain a subject (and g_3 a *with*-PP). Again, the object in g_1 prevents this group from being labelled as Instrument.

Breaking down the groups into the individual frame-argument configurations, which receive the role assignments of their groups, yields the frame-argument configuration – role type mapping displayed in table 6.2. The fact that the two object types *subj:obj/obj* and *subj:obj:pp.with/obj* are classified as Patient is straightforward. However, the unambiguous classification of the intransitive subject type *subj/subj* as Patient has only been possible by using the detour via the argument groups. The ambiguous classification of *subj:obj/subj* is appropriate, since, as discussed, the subject of "break" in a transitive sentence could express an Agent or an Instrument. In contrast, it might seem counterintuitive that the type *subj:obj:pp.with/subj* is not only labelled as Agent, but also as Instrument. One might object that a sentence expressing an Instrument by a PP cannot express another Instrument by the subject. However, the subject *can* realise an Instrument if the *with*-PP does not realise this role, e.g. in

(6.54) The hammer broke the pane with a loud bang.

Thus, this classification is appropriate. The assignment of the Instrument role to the type *subj:obj:pp.with/pp.with* is straightforward. In addition, this type is labelled as Agent because it was in a group that also contains a subject type. However, classifying a PP type as Agent contradicts the

frame-argument configuration	role type(s)
subj/subj	Patient
subj:obj/obj	Patient
subj:obj:pp.with/obj	Patient
subj:obj/subj	Agent, Instrument
subj:obj:pp.with/subj	Agent, Instrument
subj:obj:pp.with/pp.with	Instrument

Table 6.3: Refined frame-argument configuration – role type mapping for the verb "break"

linguistic observation that Agents are always expressed as subject. As emphasised in section 6.3.1, this is the only correlation between a syntactic argument and a role type that is a fact rather than a default. For this reason, a refinement of the frame-argument configuration – role type mapping is justified that excludes such cases. To achieve such a refinement, I adopted the following rule:

> **frame-argument configuration – role type mapping refinement rule:**
> Remove all assignments of the Agent role to non-subject frame-argument configurations from the retrieved frame-argument configuration – role type mapping.

Table 6.3 shows the complete mapping for "break" after applying this refinement rule.

The next step consists in dividing the items in the training data into sets so that each set corresponds to a particular thematic role type. This can be done by employing the frame-argument configuration – role type mapping in a straightforward way: For each role type, collect all data items whose left component (including the examined verb) is mapped to that role type. For illustration, let us assume that the training corpus contains the sentences (6.1)–(6.5) on page 149 and sentence (6.54) above so that the corresponding items are included in the training data. Then the item set related to the Patient comprises the following items:

```
break#subj:obj/obj  window
break#subj:obj/obj  window
break#subj:obj:pp.with/obj  window
break#subj:obj:pp.with/obj  TV
break#subj/subj  TV
break#subj:obj:pp.with/obj  pane
```

The set corresponding to the Agent contains these items:

```
break#subj:obj/subj  husband
break#subj:obj/subj  hammer
break#subj:obj:pp.with/subj  husband
break#subj:obj:pp.with/subj  wife
break#subj:obj:pp.with/subj  hammer
```

The set pertaining to the Instrument comprises the following data:

```
break#subj:obj/subj  husband
break#subj:obj/subj  hammer
break#subj:obj:pp.with/subj  husband
break#subj:obj:pp.with/pp.with  hammer
break#subj:obj:pp.with/subj  wife
break#subj:obj:pp.with/pp.with  chair
break#subj:obj:pp.with/subj  hammer
break#subj:obj:pp.with/pp.with  bang
```

These sets demonstrate that at this stage of the linking approach, the role assignment is over-generalising. A number of items associated with a role actually do not realise that role. Here, "hammer" is not an Agent, "husband", "wife", and "bang" are not Instruments. There are two reasons for this over-generalisation. Firstly, some data items are included in several sets, because their left component is mapped to multiple roles. In the example, this is the case for all transitive subject types, which are assigned the Agent and the Instrument role. Secondly, some data items are associated with a certain role according to their frame-argument configuration, but do not express that role (nor any of the other roles). Here, this applies to the last item in the Instrument set with the noun "bang". This demonstrates the necessity of semantic filters, which are intended to cope with both kinds of misclassification.

Aside of such errors, the set of items associated with a certain role provides the appropriate input to the algorithm for acquiring selectional preferences corresponding to that role. In this way, thematic role relations for WordNet can immediately be learned. For example, for learning INVOLVED_INSTRUMENT relations, the set of items corresponding to the Instrument role has to be employed. The verb–noun pairs in this set undergo the preprocessing steps described in chapter 5. First, the general mapping of verb–noun form pair frequencies to verb–noun sense pair frequencies performed as described in section 5.2 has to be applied to the items in the role-specific set. In section 5.2, I pointed out that these frequencies are computed separately for each syntactic configuration associated with a verb–noun pair in the data. For the sake of simplicity, this has not been made explicit in the corresponding equation (5.4) on page 119. However, employing the notation of verb-frame-argument configurations introduced in this chapter, this equation can be reformulated as follows:

$$freq(sns_i(v), sns_j(n)|fa(v)) = freq(fa(v), n) \times p(sns_i(v), sns_j(n)|fa(v), n) \tag{6.55}$$

With the assignment of role types to verb-frame-argument configurations, these frequencies can straightforwardly be transformed into sense pair frequencies given a certain role type:

$$freq(sns_i(v), sns_j(n)|role) = \sum_{\{fa(v)|role \in roles(fa(v))\}} freq(sns_i(v), sns_j(n)|fa(v)) \tag{6.56}$$

Equation (6.56) captures sense frequencies corresponding to those items which belong to the set associated with the role type $role$ (e.g. Instrument). The resulting counts of noun senses (depending on a certain verb and a certain role) are propagated to higher concepts in the hierarchy as

described in section 5.3. After these preprocessing steps, the learning algorithm acquires tree cut models for the verb concepts under consideration. In the example, the verb form "break" and the co-occurring noun forms ("hammer", "husband", "wife", "chair", "bang") are transformed into appropriate senses and concepts. Then, for the obtained verb concept <break>, a tree cut model is acquired from the co-occurring noun concepts obtained by the preprocessing steps.[20] The preferred concepts on that cut form candidates for establishing a role relation. For example, if the cut contains the concept <instrumentality#instrumentation> (this concept subsumes suitable senses of "hammer" and "chair"), and if this concept has a preference value above 1 (indicating preference), then this strongly suggests to adopt the relation <break> INVOLVED_INSTRUMENT <instrumentality#instrumentation>.

6.5 Semantic Filters

The last part of the linking strategy I propose is applied after learning tree cut models, i.e. it is a postprocessing step acting on the acquired tree cuts. This step is necessary to cope with the overgeneral assignment of roles to data items resulting from stages 1 and 2 of the linking strategy. We have seen that data items are associated with role types which, in fact, they do not express. The tree cut model obtained from a role-specific data item set represents these misclassified data items as well as the correct items. Continuing the example in the previous section, a tree cut model learned for the Instrument of <break> would be expected to contain adequate concepts like <instrumentality> (generalised from nouns like "hammer" and "chair") as well as inadequate concepts like <human> (generalised from nouns like "husband" and "wife") or <bang>.

The 3. linking stage has the following function: Given a tree cut which was acquired for a certain role, select those concepts on the cut which really correspond to that role and discard the other concepts. To achieve this goal, I pursue an approach which is based on a very simple idea. In short, this idea consists in associating each role type with a certain *semantic range*. Such a range is defined by a small number of rather abstract WordNet concepts. For example, we could define the semantic range of the Instrument role by the concept <inanimate_object>. That means that this range comprises that concept and all its descendants in the WordNet hierarchy. A semantic range of a certain role can be used as a filter to discard inappropriate concepts from a tree cut representing preferences for that role. To achieve that, one just has to extract that part of the cut that is dominated by the concept(s) which model(s) the semantic range of the respective role. Regarding the imaginary cut for the Instrument of <break> discussed above, this filter would approve <instrumentality>, since this is a subconcept of <inanimate_object> and thus belongs to the above-mentioned semantic range defined for Instruments. However, concepts like <human> or <bang> would be rejected, since they do not belong to that range.

Before turning to the practical issue of how to determine the semantic ranges for the respective role types, I would like to argue that the idea sketched above is not just an ad hoc strategy, but can be motivated from linguistic insights. Basically, this idea rests on the assumption that entities which fill a certain semantic role typically bear certain semantic properties which are independent of the involved verb. We have encountered such semantic characterisations of role fillers in Fillmore's deep case definitions (cf. section 2.1.1). These definitions state that an Agentive is a (typically) animate instigator, an Instrumental is an inanimate force or object, and a Locative denotes a location or spatial

[20] More precisely, a tree cut model is acquired for each verb concept the disambiguation procedure has found to represent "break".

orientation.

There is a fundamental relationship between such semantic characterisations for role fillers and selectional restrictions (or preferences) of verbs. Essentially, the semantic properties associated to a specific role can be regarded as verb-independent semantic restrictions / preferences imposed on the fillers of that role. In analogy to Jackendoff's view of selectional restrictions as the information that a verb provides about its arguments (cf. section 2.2.2), one can state that semantic restrictions on role types comprise the information that the respective roles provide about their fillers. Moreover, the semantic restrictions imposed on a role have to be compatible with, i.e. subsume, selectional restrictions that any concrete verb imposes for that role. For instance, "drive" selects a vehicle as its Instrument, "paste" glue, and "break" some solid, movable object. All these selectional restrictions are subsumed by the general characterisation 'inanimate object'. As in our case selectional preferences are represented by WordNet noun concepts, semantic characterisations of roles should be represented by WordNet concepts as well. These concepts should subsume those concepts that verbs typically prefer for the respective role, excluding concepts that typically are not preferred for that role. For example, following the examples just sketched, <inanimate_object> would be an appropriate concept for modelling the semantic preferences on the Instrument role. In a sense, a concept or a set of concepts that represents role-specific, but verb-independent preferences defines a semantic range of a role.

In section 2.1, we have seen that there is neither a consensus about the adequate inventory of roles nor on the appropriate definition of a particular role. As a consequence, it is not obvious how to delineate the semantic ranges of roles. The semantic range of a role relies on the definition of that role, which in turn is dependent on the complete role inventory. For example, consider the question whether an Agent is normally animate[21] or whether inanimate entities are also regarded as common instances of the Agent role. In Jackendoff's theory, the Instrument is analysed as being both Actor (Agent in terms of EWN) and Patient. If, following this analysis, one decides to abolish the Instrument type from the role inventory and to capture "instrumental" arguments by the Agent role, then inanimate entities may naturally serve as Agent. Anyway, independently of whether the Instrument is included in or excluded from the role inventory, Jackendoff explicitly allows for inanimate Actors, e.g. "the ball" in "The ball rolled down the hill." This is opposed to Fillmore's characterisation of the Agentive case. Hence, the question can be (and has been) decided in different ways.

For the task addressed here, it is necessary to delimit semantic ranges for the role types in EWN. The EWN role definitions are not helpful in this case, because they hardly provide any semantic restrictions on the noun concepts involved in the respective role relations. For example, the AGENT definition is: "(A/An) X is the one/that who/which does the Y." This definition characterises the involved noun by "the one/that", which is not restrictive at all. Fortunately, there is another source of information that allows to determine role-specific semantic ranges: the gold standard. For each of the roles I examine, the gold standard comprises a considerable number of relations covering a variety of verbs. Consequently, these data provide empirical evidence which is suitable to determine semantic ranges of roles. More precisely, the task of fixing these ranges amounts to the following: for each role type, find a noun concept (or a small set of noun concepts) which subsumes those noun concepts which occur in the corresponding role relations in the gold standard.

One major difficulty concerning a proper definition of the semantic range of a role is the appropriate generalisation level of the concept(s) used for that task. For example, which concept provides the most adequate characterisation of Agents: <human>, <life_form>, or <entity>? Regarding this issue,

[21] Here, "normally" means that exceptional cases of inanimate Agents, e.g. within unusual contexts or metaphoric expressions, never can be excluded.

root node	Agent		Patient		Instrument		Location	
<entity>	2188	(97.1%)	759	(74.6%)	1853	(87.9%)	158	(71.8%)
<psychological_feature>	13	(0.6%)	28	(2.8%)	25	(1.2%)	1	(0.5%)
<abstraction>	10	(0.4%)	72	(7.1%)	110	(5.2%)	4	(1.8%)
<location>	1	(0.0%)	23	(2.3%)	14	(0.7%)	54	(24.5%)
<shape#form>	0	(0.0%)	8	(0.8%)	18	(0.9%)	0	(0.0%)
<state>	7	(0.3%)	3	(0.3%)	1	(0.0%)	1	(0.5%)
<event>	3	(0.1%)	5	(0.5%)	6	(0.3%)	0	(0.0%)
<act#human_action>	7	(0.3%)	19	(1.9%)	45	(2.1%)	2	(0.9%)
<group#grouping>	25	(1.1%)	30	(2.9%)	14	(0.7%)	0	(0.0%)
<possession>	0	(0.0%)	50	(4.9%)	23	(1.1%)	0	(0.0%)
<phenomenon>	1	(0.0%)	20	(2.0%)	1	(0.0%)	0	(0.0%)
num. of concepts	2253		1017		2107		220	

Table 6.4: Number and percentage of (noun concepts of) role relations in the gold standard covered by each WordNet root node

there are two complementary desiderata: On the one hand, the concepts have to be general enough in order to be not biased towards arbitrary idiosyncrasies of the gold standard. Such idiosyncrasies can be introduced by errors or by singular role relations which significantly deviate from the majority of relations encoded for comparable verbs. On the other hand, the subsuming concepts have to be specific enough to distinguish role types. In other words, it is desirable that the semantic ranges of the different role types do not overlap. As noted, one major motivation for employing semantic filters is that some training data items are included in several role-specific data sets so that these items are captured by the respective tree cuts acquired for differing roles. If the semantic ranges of these roles are mutually exclusive, then the corresponding filters select (i.e. approve) role-specific tree cut sections which are mutually disjoint. In this way, it is guaranteed that finally each data item contributes evidence for at most one role rather than for several roles. In the example discussed in the previous section, the nouns "husband" and "hammer" belong to two data sets, pertaining to the Agent and the Instrument, respectively. If, for example, the Agent range were defined by <human> and the Instrument range by <instrumentality>, then "husband" would be captured by the selected portion of the cut acquired for the Agent, while "hammer" would not. For the Instrument, the reverse is true. If, however, the range for both roles were defined by <entity>, then both nouns were covered by the selected portion of both cuts. In the latter case, the linking strategy would ultimately not decide whether "husband" (or "hammer"), occurring as the subject of "break", is an Agent or an Instrument. Therefore, it is crucial that at least those roles whose data sets overlap are associated with mutually exclusive semantic ranges. Note that both of the desiderata just discussed have the consequence that the concepts which define the range of a role type will not capture all relations of that type included in the gold standard. To avoid the encoding of idiosyncrasies and to obtain mutually exclusive semantic ranges, the subsuming concepts that I will employ neglect a minor portion of relations.

To determine which concepts are most appropriate for defining semantic ranges of the four role types I investigate, I collected statistics about the noun concepts involved in the gold standard relations of each of these role types. In particular, I took into account different rather general noun concepts which are plausible candidates for defining role-specific semantic ranges. For these general concepts,

concept	subsumed concepts	
<entity>	2188	(97.1%)
<life_form>	2078	(92.2%)
<causal_agent>	2016	(89.5%)
<person>	2005	(89.0%)
<inanimate_object>	97	(4.3%)
<artefact>	77	(3.4%)
<group>	25	(1.1%)
<social_group>	24	(1.1%)

Table 6.5: Concepts covering a significant number / percentage of AGENT relations

I computed the amount of gold standard relations whose noun concepts they subsume, and manually selected the most adequate concepts. To get a broad overview, I started from the root concepts of WordNet. Table 6.4 shows the number and percentage of noun concepts in the gold standard that are subsumed by each root separately for each role. Obviously, this most general level is not distinctive enough, because for all roles, the vast majority of concepts is subsumed by <entity>. This means that more specific concepts have to be chosen to define semantic ranges. Therefore, I went down the WordNet hierarchy to assess which concepts are appropriate for a certain role, starting from those concepts which cover the most relations. In the following, I provide role-specific lists of concepts on different hierarchy levels which capture significant portions of relations and comment on them. In these listings, indentation indicates the levels of the concepts, relative to their ancestors and descendants. Child concepts are listed immediately below their parent, with a higher indentation.

Table 6.5 lists general concepts covering significant portions of Agent relations. More than 97% of these relations are covered by <entity>. On the level below that root node, a comparably clear tendency can be stated: Over 90% of the Agent relations in the gold standard involve living entities, mostly persons. Only about 4% pertain to inanimate objects. Concerning the question addressed above whether inanimate entities should be included in the semantic range for the Agent, these numbers give a definite answer: In EuroWordNet, inanimate Agents are marginal so that it is adequate to restrict the semantic range to living entities. The root concept with the second highest percentage of subsuming concepts is <group>, which is due to its hyponym <social_group>. This is an intuitive finding, since groups of people naturally occur as Agent. It is surprising that only 1% of Agent relations are covered by that concept.

Table 6.6 displays the concepts covering significant portions of Patient relations. Here, the picture is far less clear-cut than for the Agent role. <entity> is by far the dominant root concept, but other WordNet roots play a considerable role as well, such as <abstraction> (mostly due to <communication>, which subsumes concepts related to messages or documents) or <possession>. In general, table 6.4 shows that, as opposed to the other roles, every root node subsumes several concepts involved in Patient relations. Below <entity>, the distribution of concepts is much more heterogeneous than for the other roles. 55% of the Patient concepts are inanimate, but also a non-negligible rate of about 17% denote living entities. In sum, one can state that these findings do not justify to impose any restriction on the semantic range of the Patient.

Table 6.7 displays the concepts covering significant portions of Instrument relations. This role ex-

concept	subsumed concepts	
<entity>	759	(74.6%)
<inanimate_object>	559	(55.0%)
<artefact>	265	(26.1%)
<substance>	202	(19.9%)
<life_form>	170	(16.7%)
<causal_agent>	119	(11.7%)
<person>	117	(11.5%)
<abstraction>	72	(7.1%)
<relation>	65	(6.4%)
<social_relation>	58	(5.7%)
<communication>	58	(5.7%)
<possession>	50	(4.9%)

Table 6.6: Concepts covering a significant number / percentage of PATIENT relations

concept	subsumed concepts	
<entity>	1853	(87.9%)
<inanimate_object>	1739	(82.5%)
<artefact>	1450	(68.8%)
<instrumentality>	1071	(50.8%)
<substance>	188	(8.9%)
<material>	87	(4.1%)
<goods>	85	(4.0%)
<natural_object>	30	(1.4%)
<part>	41	(1.9%)
<body_part>	41	(1.9%)
<life_form>	65	(3.1%)
<person>	48	(2.3%)
<abstraction>	110	(5.2%)
<relation>	91	(4.3%)
<social_relation>	78	(3.7%)
<communication>	78	(3.7%)
<act#human_action>	45	(2.1%)

Table 6.7: Concepts covering a significant number / percentage of INSTRUMENT relations

concept	subsumed concepts	
<entity>	158	(71.8%)
<inanimate_object>	155	(70.5%)
<artefact>	115	(52.3%)
<construction#structure>	68	(30.9%)
<building>	26	(11.8%)
<way>	15	(6.8%)
<instrumentality>	15	(6.8%)
<substance>	21	(9.5%)
<natural_object>	9	(4.1%)
<body_of_water>	8	(3.6%)
<location>	54	(24.5%)

Table 6.8: Concepts covering a significant number / percentage of LOCATION relations

hibits a semantic profile that is almost reverse to the one corresponding to the Agent. Among the 88% portion of concepts captured by <entity>, the overwhelming majority (82.5%) is covered by <inanimate_object>, in particular by <artefact>, but also by <substance> and others. Concepts subsumed by <life_form> play a marginal role (about 3%). This justifies to exclude living entities from the semantic range of the Instrument, which avoids an intersection with the range of the Agent role. Another hyponym of <entity> which occurs in the Instrument profile is <body_part> (a subconcept of <part>). Although the percentage (scarcely 2%) is lower than the rate of <life_form>, body parts are a typical kind of Instrument. Therefore, it is adequate to include body parts in the Instrument range, the more so as this does not cause an overlap with the Agent range. Apart from <entity>, <abstraction> subsumes a noteworthy amount of relations (about 5%). This is in accordance with statements about this role in the literature (e.g. (Gruber 1965, p. 171) discusses abstract Instruments). Another root node which also represents a kind of abstract Instruments occurs with about 2%: <human_action>.

Finally, table 6.8 lists the concepts covering significant portions of Location relations. For this role type, two root concepts are of importance: <entity> (covering more than 70% of relations) and <location> (covering about 24%). At first glance, it might be surprising that <entity> is the dominant root concept rather than <location>. However, looking more closely at deeper levels of the hierarchy, this becomes more intuitive. Almost all concepts subsumed by <entity> are captured by <inanimate_object>. Most of these (52%) are artefacts, in particular pertaining to <construction> (this concept comprises e.g. buildings), but also to <way> (in the 'path' sense) or even <instrumentality> (subsuming concepts like <bed>, <camp_bed>, or <frying_pan>, which are encoded as Locations in the gold standard). Apart from <artefact>, a couple of other concepts denoting some kind of points or areas found in nature are of importance: <substance> (comprising e.g. liquid, water, or ice), <natural_object>, and <body_of_water>. All these concepts may naturally be referred to as Location, although they are not subsumed by <location>. Overall, this semantic profile is clearly distinct from the Agent profile. In contrast, it cannot be so clearly distinguished from the Instrument profile. Both exhibit a strong focus on inanimate objects. Although within this subrange there is emphasis on different kinds of concepts, there is a considerable overlap even on deeper hierarchical levels (e.g. <instrumentality> or <substance>). Therefore, it is not possible to define semantic ranges for these two role types which do not intersect.

Based on these profiles, I fixed the following general concepts to define semantic ranges for the different roles:

Agent: <life_form>, <causal_agent>, <group#grouping>

Patient: —

Instrument: <inanimate_object>, <body_part>, <abstraction>, <act#human_action>

Location: <inanimate_object>, <location>

These ranges reflect the two desiderata mentioned above. On the one hand, the concepts I selected are at a rather general level, e.g. <inanimate_object> instead of <artefact>, <building>, <way>, or <body_of_water>. In this way, the range definitions abstract from idiosyncrasies in the gold standard. For the same reason, I did not take into account root concepts which only represent a marginal percentage of relations for a particular role (e.g. <psychological_feature> for all role types). Exceptions are the inclusion of <group#grouping> in the Agent range as well as <body_part> and <human_action> in the Instrument range. These concepts denote typical instantiations of the respective roles, although they are under-represented in the gold standard.[22] On the other hand, the selected concepts are specific enough to allow distinctions between roles where possible. In particular, the root node <entity>, which is predominating for all roles, would be too general. Restricting the Agent range to living entities and the Instrument and Location ranges to inanimate entities allows to distinguish Agents from Instruments or Locations.

In contrast, the examination of the gold standard shows that it is not possible to fix semantic ranges for Instrument and Location that are mutually exclusive and thus allow to discriminate these role types by semantic filters. Therefore, I had to introduce linking heuristics which do not concurrently assign both role types. I discussed this issue in section 6.3.4. I formulated the Instrument heuristic in a way that it does not assign the Instrument label to an argument group which is classified as Location. Hence, the linking heuristics already do the job of distinguishing these two role types. The situation is even worse for the Patient. As table 6.4 shows, about 75% of Patient relations is captured by <entity>. The rest is more or less equably divided among the other root nodes. Discarding these nodes and defining the semantic range of the Patient by <entity> would have the consequence that about one quarter of the relations would not be taken into account. Moreover, a further restriction of the <entity> range would be inadequate, since inanimate objects as well as life forms are represented with a significant rate. Hence, the Patient role cannot be associated with a semantic range at all. For this reason, the linking rules have to separate the Patient from the remaining roles: If an argument group is labelled as Patient, then the rules ensure that this group is not at the same time labelled as another role.

I would like to conclude this section with some general remarks concerning the approach of determining semantic filters by examining the gold standard. One could object that this approach is not admissible, because it tunes the tree cuts to be evaluated by adapting them to the test data which are used for the evaluation. More specifically, first the tree cuts are filtered so that only those concepts remain which are captured by the gold standard, and after that the evaluation experiments examine to what extent these manipulated tree cuts match with the gold standard. In the following, I provide several arguments to rebut this objection.

[22] Actually, one could regard the low percentage of these concepts itself as an idiosyncrasy of the gold standard.

First of all, the investigation of the role-specific semantic profiles in the gold standard proved that these profiles are not arbitrary. In contrary, they are intuitive and, in essence, internally consistent. Furthermore, they conform to semantic characterisations of the respective roles provided in the literature. This is a remarkable finding, considering the fact that the gold standard is no homogeneous data set, but originates from several independent sources (different language-specific wordnets in EWN) and the role relations in these wordnets have not been encoded for the purpose of providing a balanced and representative test set. It is necessary to examine the semantic profiles for two reasons. Firstly, it is important to find out how the gold standard "behaves" with respect to controversial role characterisations. For instance, concerning the question whether it is adequate or not to exclude inanimate objects form the Agent range (both alternatives are linguistically justifiable), it is crucial to know to what extent inanimate objects are encoded as Agents in the gold standard. Secondly, it is not immediately clear how quite general and informal semantic characterisations provided in the linguistic literature are captured by the hierarchical structure of WordNet. For example, buildings naturally belong to the semantic range of the Location role. However, in WordNet <building> and its hyponyms are not subsumed by <location>, but by <inanimate_object>. For such reasons, the latter concept must be included in the Location range. Looking at the semantic profiles of the different roles helps to recognise such subtleties, which would be likely to be missed if role ranges were fixed by mere introspection. Nevertheless, although the profiles extracted from the gold standard guided my definitions of semantic ranges, they did not exclusively determine them. I also took concepts into account which are only marginally reflected by these profiles, but which intuitively belong to the role range in question (e.g. <group> for the Agent or <body_part> for the Instrument). Thus, the semantic ranges defined above are motivated both empirically and introspectively.

Another argument against the objection mentioned above is that one crucial aspect of the task of this thesis is to acquire selectional preferences at an appropriate level of generalisation. This challenge is in no way alleviated by restricting the acquired preferences to a limited, but still very broad range. The task is to find the appropriate abstraction level within that general range.

Finally, in the next chapter I empirically investigate the impact of semantic filters. In our setting the practical consequences of applying semantic filters to the acquired tree cuts is two-fold. On the one hand, the number of recognised relations in the gold standard may be decreased, because a minor portion of gold standard relations is not captured by the role ranges. This would result in a decrease of recall. However, this effect will be marginal, because the portion of dropped relations is small. For each role type, the range defined above captures more than 90% of the corresponding relations in the gold standard. On the other hand, since the filters discard large portions of the learned tree cuts, the number of candidate concepts for the examined role is reduced. One can expect that this reduction will be significant, which would speed up the manual inspection of the remaining candidate concepts. However, this question has to be examined empirically. Therefore, in the experiments described in section 7.4, I will test the impact of employing semantic filters by comparing recall and the number of candidate concepts with and without applying these filters.

6.6 Related Work

At the end of this chapter, I would like to briefly discuss the major commonalities and differences between the linking strategy I have proposed here and related approaches proposed in the NLP literature. It is important to note that these approaches serve different purposes than my strategy. In my work, linking serves to transform the training data, which consist of syntactic verb–noun relations, to

semantic verb–noun relations. These relations form the input to the algorithm for learning selectional preferences so that this algorithm immediately acquires thematic role relations. In contrast, the work I refer to in this section pursues different goals: a general semantic characterisation of verbal arguments within a particular domain; verb classification, especially detecting a verb's participation in diathesis alternations; automatic labelling of semantic roles in a corpus.

The work which perhaps is closest to my linking strategy has been done by Peters (1996). Peters aims at acquiring a semantic characterisation of thematic arguments of verbs within a constrained sublanguage. For his feasibility study, he uses a corpus containing texts about satellite communication. This corpus was automatically tagged and parsed. In a first step of his analysis (which is performed with four high-frequency verbs), simple rules map syntactic arguments to thematic roles. For example, the NP following the verb is classified as THEME, PPs headed by "over" or "through" as PATH, PPs headed by "from" as SOURCE, or PPs headed by "with" as INSTRUMENT / CONCOMITANT. A second step acquires coarse-grained semantic characterisations for the thematic arguments of the examined verbs. To this end, Peters defines ten abstract semantic classes which represent very general WordNet concepts, e.g. 'Inanimate object', 'Agent' (effectively comprising the concepts <causal_agent> and <life_form>), 'Activity', or 'Group, Organisation'. To retrieve the semantic characterisation for a role of a verb, each of these classes is associated with the number of different nouns (types rather than tokens) occurring at the respective argument position which are subsumed by the corresponding WordNet concepts. In effect, these semantic characterisations can be viewed as a rudimentary form of abstract selectional preferences, where the preference values consist in counts of noun types.

There are a number of similarities between this approach and my linking strategy. In particular, the rules of mapping syntactic arguments to thematic roles resemble the linking heuristics which I employ in the second stage. However, these rules immediately act on the parsed sentences rather than on argument groups. Moreover, except for the INSTRUMENT / CONCOMITANT ambiguity (see below), they establish a one-to-one mapping between syntactic arguments and thematic roles, which my heuristics do not. This unique mapping is possible due to two reasons (that are mentioned by Peters himself). Firstly, Peters does not take into account the subject of a sentence, which, as we have seen, is most ambiguous regarding the role it might express. Secondly, within a highly restricted domain (which is intended in Peters' work), the use of prepositions is likely to be restricted in a way that a certain preposition uniquely realises a certain role. In contrast, since the task of this thesis is the acquisition of roles for general-purpose lexical resources, which implies processing unrestricted text, my linking strategy has to be much more sophisticated. One exception in Peters' study is "with". This preposition is ambiguous; it might indicate an INSTRUMENT or a CONCOMITANT. Peters suggests that this ambiguity can easily be resolved by separating the abstract semantic classes according to animacy/non-animacy. In a sense, this corresponds to the semantic filters I use in stage 3. To conclude, Peters' approach shares some basic ideas with my strategy, but is very simplistic in many respects, (linking rules, preference acquisition, restriction of domain and syntactic structures).

In the past years, much work has been done to automatically classify verbs with respect to patterns of subcategorisation and selection. Essentially, this work was inspired by (Levin 1993), who did not only provide a comprehensive survey of diathesis alternations, but furthermore developed a detailed system of semantic classification of verbs which is based on the hypothesis that the semantic class a verb belongs to is strongly correlated with the set of alternations which this verb undergoes. Meanwhile, numerous publications addressed the issue of automatically obtaining such a kind of classification (including the extension of Levin's classification system and its coverage), or the related issue of detecting the participation of verbs in diathesis alternations. The information employed for these goals

consists of syntactic and semantic properties of verbs, which are extracted from corpora by means of quantitative methods. (McCarthy 2001, p. 110–125) provides an excellent overview about the work within this area. In this section, I will concentrate on two approaches, which have been proposed in (Schulte im Walde 1998*b*) and in (McCarthy 2001), respectively. These approaches are particularly interesting in our context, because each of them employs techniques which are very similar to the ones I make use of.

Schulte im Walde (1998*b*) applies clustering techniques to acquire verb classes. She uses the same training data that I use. (As noted in section 5.1, I received the data from her.) Verbs are clustered according to two kinds of information: subcategorisation frames and selectional preferences. In detail, Schulte im Walde extracts two kinds of input to the clustering mechanism from the data. The first kind consists of pairs of verbs and corresponding subcategorisation frames, e.g. *(fly, subj)* or *(fly, subj:obj)*, and probabilities associated with these pairs. The second kind is similar, but the argument slots in the frames are supplemented with selectional preferences expressed by high-level WordNet concepts, e.g. *(fly, subj(<physical_object>)* or *(fly, subj(<life_form>):obj(<physical_object>)*.[23] Clustering is performed separately for each kind of pairs, i.e. with and without taking selectional preferences into account. Furthermore, two different clustering methods are used and evaluated: an iterative clustering approach and the LSC approach, which I employ for grouping frame-argument configurations. However, note that the pairs to be clustered do not consist of a verb-frame-argument configuration and a noun, but of a verb and a frame (optionally supplemented with selectional preferences). For evaluating the resulting clusters, Levin's classification is used as a gold standard. A cluster is judged as accurate if the verbs it contains[24] form a subset of a Levin class. Surprisingly, the approach performs better without including selectional preferences. (Schulte im Walde 1998*b*) explains this finding as a problem of data sparseness, which becomes more salient for a more complex statistical model with more parameters. In more recent work, Schulte im Walde performed similar experiments for German data (cf. (Schulte im Walde 2003)). She provides a qualitative evaluation of the verb clusters retrieved by her experiments and concludes that for some verb classes, accuracy should be improved by employing more fine-grained selectional preferences, while for other verb classes, preferences at a lower abstraction level would decrease accuracy. This immediately raises the question of the (verb-dependent) appropriate generalisation level of selectional preferences, which is a key issue of this thesis. It would be interesting to investigate to what extent my approach can be employed for the work of Schulte im Walde to increase the performance of automatic verb classification.

McCarthy (2001) addresses the task of detecting verbal participation in diathesis alternations. She proposes strategies which, given a particular alternation (e.g. the causative alternation), test whether verbs that exhibit syntactic argument structures corresponding to both alternation variants really undergo this alternation or not. For example, "break" and "eat" occur as transitive or intransitive verbs, i.e. in both subcategorisation variants pertaining to the causative alternation. However, "break" participates in this alternation, whereas "eat" does not. Candidate verbs for a particular alternation are found by employing subcategorisation information which has been extracted from a parsed corpus. McCarthy develops several approaches to determine participation or non-participation of these candidates in the alternation in question. These approaches are based on the same general idea which underlies stage 1 of my linking strategy (argument grouping): She measures the similarity of the frame-argument configurations in the alternation variants which correspond to each other (e.g., for the causative alternation, the object of the transitive variant is compared with the subject of the in-

[23] The estimation of probabilities for these pairs take into account the preference values of the respective concepts, which are computed according to Ribas' approach.

[24] For each LSC cluster, only the four verbs with the highest class probabilities are taken into account.

transitive variant). To this end, she employs the tree cut approach of Li and Abe to acquire a tree cut model for each of the frame-argument configurations to be compared. She investigates several methods to assess the similarity of these tree cut models. If they are judged to be similar, then the verb is classified as undergoing the alternation, otherwise it is not. My strategy measures the similarity of frame-argument configurations by employing a LSC model rather than tree cut models. Both methods have in common that the models which represent the semantic profiles of the frame-argument configurations perform a kind of smoothing over the nouns occurring as the respective arguments. This overcomes data sparseness problems. McCarthy also tests a method that measures similarity by just comparing the bare multisets of nouns which occur as fillers of either frame-argument configuration; here, the portion of overlap of the two multisets quantifies the frame-argument configurations' similarity. McCarthy's experiments show that this method performs worse than the approaches employing tree cut models.

Note that there is a major conceptual difference between the setting of McCarthy's approach and the setting of my linking strategy. McCarthy starts from concrete diathesis alternations and tests the similarity of frame-argument configurations to judge whether particular verbs take part in these alternations. In contrast, I first establish groups of similar frame-argument configurations regardless of specific diathesis alternations, and in the following step, I apply role labelling heuristics which to a large degree are based on knowledge about such alternations. However, my grouping strategy can straightforwardly be adapted to fit in McCarthy's setting: given a particular alternation, a verb under examination undergoes this alternation if the corresponding frame-argument configurations are grouped together. It would be interesting to test the performance of the argument grouping strategy for that task.

Finally, the work initiated by (Gildea & Jurafsky 2002) is highly relevant in our context. Gildea and Jurafsky develop a supervised approach for automatically labelling semantic roles in corpus sentences. This approach uses a training corpus which has been manually annotated with semantic role markings. This corpus has been compiled within the FrameNet project. For each sentence, the annotation specifies the so-called *target word*, i.e. a predicate (in most cases a verb) which is associated with a certain conceptual frame, as well as those constituents that express particular roles of that frame. For example, the JUDGEMENT frame comprises the roles JUDGE, EVALUEE, and REASON. "blame" is a verb that invokes this frame. A sentence containing that verb is annotated as follows:

(6.57) [*Judge* She] **blames** [*Evaluee* the Government] [*Reason* for failing to do enough to help].

The example illustrates that the role inventory used here is less general than the thematic role types which I deal with in this thesis. (However, Gildea and Jurafsky address the task of learning abstract thematic roles by mapping the FrameNet role types to corresponding roles like Agent, Patient, Path, or Proposition.) The sentences annotated in this way are automatically parsed and then employed for training a statistical classifier (i.e. a statistical model) to be used for role labelling. This model incorporates probabilities of specific features given a particular role type, which are estimated from the training data. These features include the target word invoking the frame, the voice of the target word, the phrase type (NP, PP, etc.) of the constituent expressing the role in question, the grammatical function of that constituent, its location in the parse tree relative to the target word, and its head word. The statistical classifier combines these probabilities by a backing-off mechanism and predicts the appropriate role expressions for unseen sentences. Gildea and Jurafsky investigate two variants of the labelling task. In the easier variant, the boundaries of the role expressions (i.e. the constituents which actually express some role) are known and just have to be labelled correctly. In the more difficult

variant, these boundaries have to be predicted as well. Strategies of generalising from the head words to overcome the sparse data problem are also investigated, among them LSC clustering and the use of WordNet. These techniques prove successful for increasing the performance of the system.

Recently, the work of Gildea and Jurafsky has inspired the setup of closely related *shared tasks* (competitions) of semantic role labelling in association with two different conferences, namely Senseval-3 and CoNLL-2004. These shared tasks in turn have initiated various research efforts within that field. For an overview of the respective task definitions as well as the submitted contributions, cf. (Litkowski 2004) for Senseval-3 and (Carreras & Màrques 2004) for CoNLL-2004. For both tasks, training data had been provided comprising a set of parsed sentences[25] in which target words were marked and the constituents realising associated semantic roles were labelled accordingly. For the Senseval-3 task, these sentences had been taken from the FrameNet corpus; for the CoNLL-2004 task, the data had been extracted from PropBank (cf. (Kingsbury, Palmer & Marcus 2002)), a corpus of syntactic trees (originating from the PennTreebank) supplemented with information about predicate-argument relations, in particular semantic role identifiers. In both settings, the task was to employ these data for learning a classifier which assigns semantic role labels to unseen test data. For these shared tasks, a number of contributions were submitted, applying various machine learning approaches such as clustering techniques, Brill's Transformation-based Error-driven Learning, Maximum Entropy, Memory-Based Learning, Support Vector Machines, etc.

In a sense, the task addressed by (Gildea & Jurafsky 2002) and subsequent work is at the same time more and less ambitious than my linking strategy and its utilisation for the task of this thesis. Labelling semantic roles in sentences is equivalent to retrieving a unique assignment of semantic role types to role-expressing constituents. In contrast, as we have seen in section 6.4, stages 1 and 2 of my approach in general yield a non-unique mapping between verb-frame-argument configurations and roles, i.e. a verb-frame-argument configuration (and thus all corresponding data items) might be associated with multiple role types. The design of the semantic filters in stage 3 aims at indirectly compensating this shortcoming: As the semantic ranges of those role types which might concurrently be assigned to an argument group are mutually exclusive, each data item should contribute to only one filtered role preference model, provided that argument grouping and word sense disambiguation work correctly. Apart from that, the information learned by the method of Gildea and Jurafsky and related approaches is more sophisticated than the information acquired by my strategy w.r.t. a further aspect. While my approach only takes head nouns of verb complements into account, the setting put forward by Gildea and Jurafsky also includes (in its more difficult variant) the identification of the boundaries of role expressions, as well as the recognition of roles which are realised as adjuncts or subconstituents of complements. As the task of this thesis consists in acquiring models of role realisations which generalise from concrete utterances, it is justifiable that the information I deal with abstracts from various aspects of individual realisations of roles.

On the other hand, the sophisticated task of semantic role labelling requires supervised learning methods, at least at the current state of the art. This, in turn, requires training corpora which include annotated role labels. To my knowledge, such data are currently available only for English, as part of the FrameNet database and in PropBank. FrameNet-like resources are under construction for a few other languages (German, Spanish, Japanese). In contrast, my linking strategy is an unsupervised approach. It solely requires resources which are available for a considerable and yet increasing number of languages. In this context, it is worth noting that this strategy is hybrid, i.e. it integrates a quantitative clustering approach (stage 1) and rules based on linguistic and conceptual knowledge (stage 2 and 3),

[25]The Senseval-3 data contained complete parse trees, whereas the CoNLL-2004 data only contained partial parses (chunk parses).

while the classifiers developed by Gildea and Jurafsky and within the work submitted for the shared tasks of Senseval and CoNLL in general do not incorporate the latter kind of knowledge (apart from the selection and combination of features). Apparently, linguistic regularities which are stipulated in my approach are implicitly available in and learned from the role-labelled training data. In summary, the information effectively acquired by my linking strategy is poorer than the information acquired for a role labelling system. However, it is sufficient for learning thematic role relations. The unsupervised method to acquire it is in principle applicable for numerous languages. The necessary adaptation of the linguistically motivated rules to a specific language requires much less effort than creating suitable corpora for supervised training of a linking system. In case such data are available, they are very useful for the task of learning role relations as well: Corpora which are manually or automatically role-labelled could immediately be used as input for algorithms learning selectional preferences. But since the task of this thesis includes the maximal adaptability to different languages, an approach that does not need such corpora is preferable, at least until they are as widespread across languages as wordnets themselves.

Chapter 7

A Detailed Evaluation

This chapter provides an in-depth evaluation of the approach for learning thematic role relations I developed in this thesis. This evaluation investigates the performance of acquiring relations of the four role types which I selected for that task and which I elaborately discussed in the previous chapter, i.e. AGENT, PATIENT, INSTRUMENT, and (to a limited extent) LOCATION. To measure this performance, I employ the gold standard introduced in section 5.4. Recall that this gold standard, although suitable in general, exhibits some inadequacies with regard to the evaluation task. In section 7.1, I briefly repeat these inadequacies and describe some simple heuristics I applied to overcome them to a certain degree. Nonetheless, in the following sections we will encounter other shortcomings of the gold standard which significantly affect the evaluation results. Section 7.2 outlines the general setup of the evaluation experiments and states the evaluation criteria. After that, I turn to the experiments themselves: In section 7.3, I report the results of basic performance tests. The subsequent sections describe elaborations and modifications of these tests, which focus on how different factors influence performance. First, I investigate the impact of the linking strategy proposed in chapter 6, in particular the application of semantic filters (section 7.4) and the clustering approach for mapping syntactic arguments to thematic relations (section 7.5). Other factors I examine comprise the parameters of the LSC model employed by the learning approach (section 7.6), the treatment of virtual leaves in the concept hierarchy (section 7.7), and the generalisation level of the noun concepts in the gold standard (section 7.8). Section 7.9 summarises the main findings of the experiments. In general, it will turn out that these findings do not only concern the performance of the learning algorithm, but at the same time provide substantial insights into certain properties of the gold standard and its suitability for the evaluation task.

7.1 The Gold Standard

In section 5.4, I sketched how I extracted a gold standard of thematic role relations from the EuroWordNet database: Several language-specific wordnets in EWN contain such relations. Replacing the verb and noun concepts involved in these relations by the corresponding concepts in the Interlingual Index (which essentially consists of all WordNet concepts) yields a set of role relations for WordNet. As discussed in section 5.4, this gold standard has some considerable weaknesses. Generally, the translation step from the individual wordnets to WordNet may introduce some inaccuracies due to deviating selectional preferences of corresponding verbs in different languages, differing hi-

erarchical structures of the language-specific wordnets, or plain errors, which usually occur in any manually built resource. However, having inspected a considerable portion of the gold standard (especially the PATIENT relations, cf. section 5.5.1), I regard the actual amount of such inaccuracies as tolerably low. A more serious problem concerning the evaluation task is caused by the intended semantic nature of part of the role relations encoded in EWN. Some relations capture incorporated arguments of a verb rather than arguments which are realised as syntactic complements, e.g.

(7.1)

a. <delouse> INVOLVED_PATIENT <louse#sucking_louse>

b. <silt_up#silt> INVOLVED_PATIENT <deposit#sedimentation#alluvium>

Obviously, the approach proposed in this thesis cannot detect incorporated arguments, since they are not visible in corpus data. Another kind of relations includes noun concepts which provide an intensional characterisation of a preferred argument rather than a generalisation of the preferred extensional range, e.g.

(7.2) <address#speak_to#turn_to> INVOLVED_PATIENT <addressee>

In this example, <addressee> adequately characterises the preferences of <address> for its Patient. However, it is unlikely that such intensional characteristics will be captured by an approach that generalises from corpus instances. In this case, <addressee> is a hyponym of <person> and does not have any hyponyms itself. In the training corpus, however, the objects of "address" (which are recognised as Patient by the linking strategy) denote, among others, several kinds of person (e.g. "captain", "customer", or "member"), but not "addressee". Thus, the learning strategy pursued in this work would probably acquire <person> as the appropriate generalisation of these instances, but in any case miss <addressee>, since the corpus instances are not subsumed by that concept in WordNet.

To allow for a fair evaluation, relations like in (7.1) and (7.2) should be excluded. As a manual inspection of the complete gold standard to eliminate relations would be too time-consuming and error-prone, I decided to apply a simple heuristic to filter unsuitable relations automatically. This heuristic rests on the reasonable assumption that role relations which encode incorporations or intensional preferences tend to involve nouns and verbs which are derivationally related to each other. Note that this assumption holds for (7.1 a.) and (7.2), though not for (7.1 b.). I employed a straightforward method to automatically recognise derivations: For every relation, each verb form in the verb synset is compared to each noun form in the noun synset. If a verb string is completely contained in a noun string or vice-versa, then the relation is regarded as derivational. This strategy recognises the derivational relationships in (7.1 a.) ("louse" is contained in "delouse") and (7.2) ("address" is contained in "addressee"). However, it misses some obvious cases of derivation, e.g. in

(7.3) <produce#bring_about#give_rise_to> INVOLVED_PATIENT <product#production>

For this reason, I refined the heuristic as follows: If the comparison of verb and noun forms do not detect derivation, then delete a trailing 'e' from each verb form if present. If a verb string modified in this way is contained in a noun string, then the relation is regarded as derivational. This also captures (7.3), since the shortened verb string "produc" is contained in "product".

role type	#
AGENT	1685
PATIENT	888
INSTRUMENT	1621
LOCATION	173

Table 7.1: Number of different types of role relations in the gold standard after filtering

If a relation exhibits a derivation, then it is not necessarily adequate to eliminate it. In some cases, a role relation involves a noun concept that is an intensional characterisation *as well as* an extensional generalisation of the preferences of the involved verb concept. For example, the noun concept <product#production> in (7.3) subsumes a number of concepts in the WordNet hierarchy, e.g. <book>, <handcraft>, or <software>. Such relations should not be excluded, because here the learning algorithm has a "fair chance" to find the noun concept by generalising from corpus instances. Therefore, I applied the following heuristics to eliminate relations from the gold standard: If a relation is classified as derivational, then it is deleted if the noun concept has less than 5 subconcepts (i.e. direct or indirect hyponyms). This strategy is very coarse-grained and does not capture all unsuitable relations (e.g. (7.1 b.)). However, the portion of such relations is reduced significantly. Table 7.1 shows the number of relations in the gold standard for each examined role type after the filtering step. A comparison with the original numbers (cf. table 5.1 on page 140) shows that filtering eliminated about 25% of AGENT relations, 13% of PATIENT relations, 23% of INSTRUMENT relations, and 21% of LOCATION relations.

7.2 Experimental Setup, Evaluation Criteria, and Parameters

The evaluation experiments described in this chapter have a very similar setup to the experiments reported in section 5.5. The main difference is that the experiments in the following sections take into account several role types and include the linking strategy developed in chapter 6 (apart from section 7.4 and 7.5), whereas the tests in chapter 5 are restricted to the PATIENT relation and employ the simple linking heuristic that all and only syntactic objects express Patients. All other processing steps remain the same. The training data described in section 5.1 are used. In a preprocessing phase (following stages 1 and 2 of the linking strategy), word forms are related to WordNet concepts via the frequency count technique sketched in section 5.3, optionally preceded by the word sense disambiguation strategy presented in section 5.2. The acquired tree cuts are compared to the gold standard, separately for each role type. To avoid data sparseness problems, I only take into account verb concepts with a frequency of at least 50.

The usual measures to assess the performance of an NLP technique are precision and recall. In our case, recall corresponds to the percentage of relations in the gold standard which are acquired by the learning algorithm, while precision corresponds to the percentage of acquired relations which occur in the gold standard. I pointed out already in section 5.5 that, while it is no problem to employ the recall measure, it would be inappropriate to use the precision measure. The reason for this is that the gold standard does not necessarily provide an *exhaustive* encoding of the selectional preferences of a verb concept. Thus, it is common that the learning algorithm acquires preferred concepts for a verb

which are perfectly adequate, but which are not covered by the gold standard. Employing precision for evaluation would punish this behaviour, because it would treat such correctly acquired preferences as errors. For this reason, the primary measure I adopt to test performance is based on recall. However, the measure I employ deviates from the above-mentioned definition of recall in one important respect: It is not based upon *all* relations in the gold standard, but only on the relations with a verb concept for which the algorithm effectively retrieves role relations of the examined type. Actually, it happens that for some verb concepts in the gold standard no relation is learned at all. The reason for this might be found in any stage of the acquisition process, e.g. the corresponding verb forms do not occur in the data with sufficient frequency (or not at all), or the correct verb senses are erroneously "disambiguated away", or the semantic filter eliminates all noun concepts which are preferred by the verb concept. These factors relate to the performance of the learning approach and its evaluation in very different ways. For example, the evaluation should not penalise a verb's absence in the data, whereas the appropriate construction of semantic filters should be subject to evaluation. Unfortunately, however, it is not trivial (if possible at all) to determine the "responsible factor" in each individual case. Thus, I decided to leave those verb concepts aside for which no relations are learned. In other words, the evaluation presented here focuses on the question: *If the learning approach proposes some relations for a verb concept, how good are these proposals?* The measure I use does not depend on the absolute coverage of the gold standard by the acquired results. Recall as defined above would include that aspect. Therefore, to avoid confusion, I will use the term *accuracy* instead of recall for the primary performance measure. I will always explicitly state the respective number of gold standard relations taken into account in each experiment. This information will be of interest in some cases.

As in section 5.5, accuracy is differentiated as to how precisely a relation learned by the acquisition algorithm matches a gold standard relation. In particular, the evaluation captures the number and the percentage of the noun concepts in gold standard relations which were exactly matched, not matched at all, or matched by more general or more specific concepts in the tree cut models retrieved for the respective verb concepts. This makes it possible to measure to what extent the cuts exactly agree with the gold standard, miss it completely, or exhibit a level of generalisation which is too high or too low. Note that I will consider approximate matches, i.e. concepts on the cut which match the gold standard with a deviation of at most one hierarchical level, as sufficiently adequate. This is justifiable because human intuition might slightly vary w.r.t. the *exact* generalisation level of a verb's selectional preferences. Even the gold standard itself contains cases of one-level variations, e.g. the two INVOLVED_PATIENT relations (originating from different wordnets) from <own#have#possess> to <asset> and <possession>, respectively (cf. section 5.4). In this context, it is worth noting that the average depth of the WordNet 1.5 noun hierarchy amounts to 7.2. The maximal hierarchy depth is 15.

Despite the above-mentioned drawbacks of the precision measure in our context, it is highly desirable to include an evaluation criterion that complements our recall-related accuracy measure in some way. If the evaluation only concentrated on recall, it would be biased to favour tree cut models with a high number of preferred concepts. This is because a tree cut that the algorithm learns for a certain verb (provided it is located at the appropriate generalisation level) is more likely to match the noun concept(s) listed for that verb in the gold standard if it comprises a large amount of preferred concepts. On the other hand, however, the more concepts a cut contains, the more concepts could be erroneously acquired. Recall that one of the basic application scenarios assumed for the learning approach developed in this thesis is the semi-automatic acquisition of role relations: Given a verb concept and a role relation type, the algorithm acquires a set of corresponding noun concepts, which are manually inspected afterwards. Within this scenario, it is preferable that the set of acquired noun concepts is rather small so that the effort of manual examination is minimised. Therefore, to quantify this manual effort,

I decided to adopt the average number of preferred noun concepts per verb, i.e. the average number of relations (of a specific role type) the algorithm learns per verb, as a second evaluation criterion. Note that this quantity is related to precision. The total number of acquired relations corresponds to the denominator of the precision ratio. The numerator, i.e. the number of correctly acquired relations, cannot be determined automatically (s.a.). Hence, although it is not possible to employ the precision measure, at least its known component is taken into account. I decided to employ the average number of acquired relations per verb rather than the total number of acquired relations. This is motivated by two reasons: firstly, the average number of relations per verb is more meaningful w.r.t. the effort of manual postprocessing; secondly, since the number of verbs captured by the gold standard differs significantly for different role types, the only way to compare the amount of relations across role types is to average this amount over the involved verbs. Hence, I use two measures as evaluation criteria: accuracy (a modification of recall) and the number of acquired relations per verb. There is no obvious way of combining these two quantities as, for example, the F-measure combines precision and recall. Therefore, these two quantities have to be treated as independent characteristics.

Independently from the different aspects I examine in turn in the subsequent sections, there are two general parameters which have a major impact on the evaluation results. As in section 5.5, I examine this impact by using different settings in the experiments described below. The first of these parameters is the constant C, which influences the degree of generalisation. I will test several values of this constant in order to optimise performance. The second parameter is the alternative to employ or to skip word sense disambiguation when preprocessing the data. The experiments in section 5.5 suggest that WSD might have a contradictory effect on our two evaluation measures, which illustrates the trade-off between them. On the one hand, WSD reduces the number of acquired relations, since disambiguation abolishes most of the potential word senses; therefore many verb and noun concepts are "disambiguated away" so that the algorithm does not acquire relations between them. This is the desired positive effect. On the other hand, as the WSD approach (like any of the current WSD techniques) is not free of errors, a certain amount of correct senses are dropped as well. This has a negative effect on accuracy, because the more correct senses are missing, the more concepts which match the gold standard get lost.[1] The better the WSD technique, the more this negative effect is reduced, while maintaining the positive effect of sorting out erroneous candidate relations. Therefore, taking into account results for both disambiguated and non-disambiguated data gives a rough idea of the potential of the learning algorithm: An optimised WSD method should approximate the accuracy achieved with non-disambiguated data and the number of relations per verb acquired with disambiguated data.

7.3 Basic Performance

This section reports the results of experiments with the basic settings as outlined above. In particular, the series of tests described here employ different values of C, namely 100, 1000, 10 000, 1 000 000, and 1 000 000 000.

[1] Note that this is an empirical finding. The sketched interrelation is highly plausible, but not necessary. It could also happen that concept frequencies and probabilities estimated from non-disambiguated data are negatively biased by erroneous senses so that for many concepts matching the gold standard the preference value would indicate dispreference. In that case, recall would be low for non-disambiguated data. However, the experiments in chapter 5 show that recall is higher without WSD than with WSD. Obviously, the probabilities of correct concepts overall are still high enough to let them be classified as preferred, but a high number of erroneous concepts receive high preference values as well.

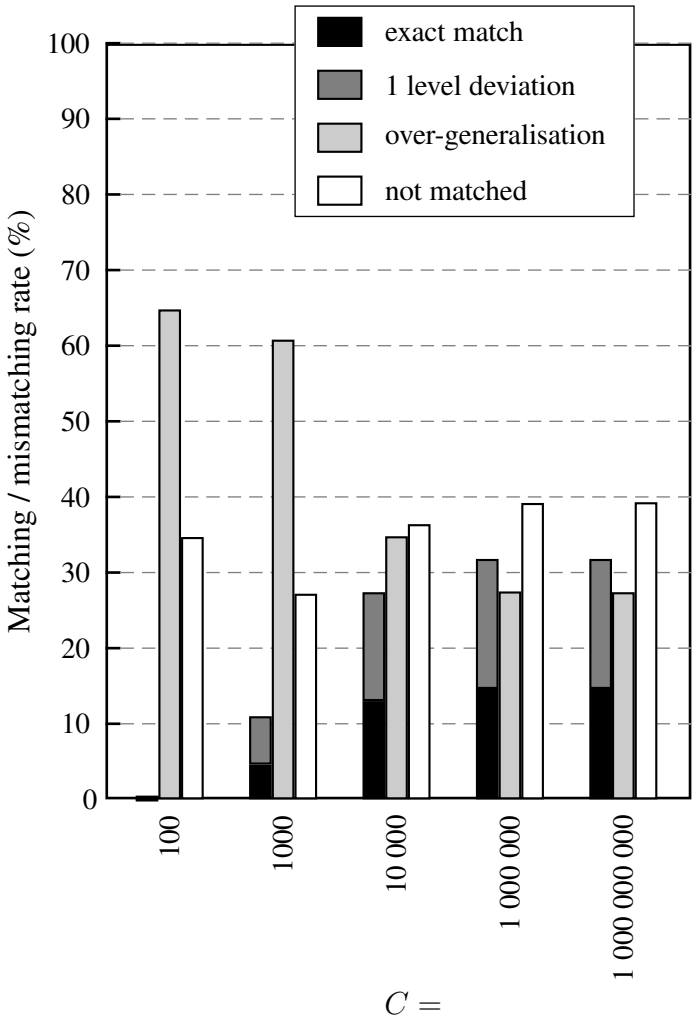


Figure 7.1: Gold standard matching rates for AGENT applying different values of C with WSD

7.3.1 AGENT

Tables 7.2 to 7.6 show the results for the AGENT role. Figures 7.1 and 7.2 provide bar diagrams which visualise the most interesting matching and mismatching rates with and without employing WSD, respectively. In general, with growing value of C accuracy increases as well. The maximal rate of approximate matches (matches with a deviation of 0 or 1 levels in the hierarchy) is 40%, accomplished with $C = 1\,000\,000\,000$ and without WSD. This value is far below the results achieved in the experiments reported in section 5.5. The minimal percentage of relations not matched at all by the acquired cuts is achieved with $C = 1000$, where most concepts are matched by over-general concepts. With C growing above that value, the portion of relations captured by over-generalisation decreases, with the side effect that more concepts are completely missed.

Apart from a rather high percentage of gold standard concepts which are not matched at all, a large amount of concepts is matched by too general concepts on the acquired tree cuts. This would suggest that the abstraction level of the cuts is too high. However, even for very high values of C, about

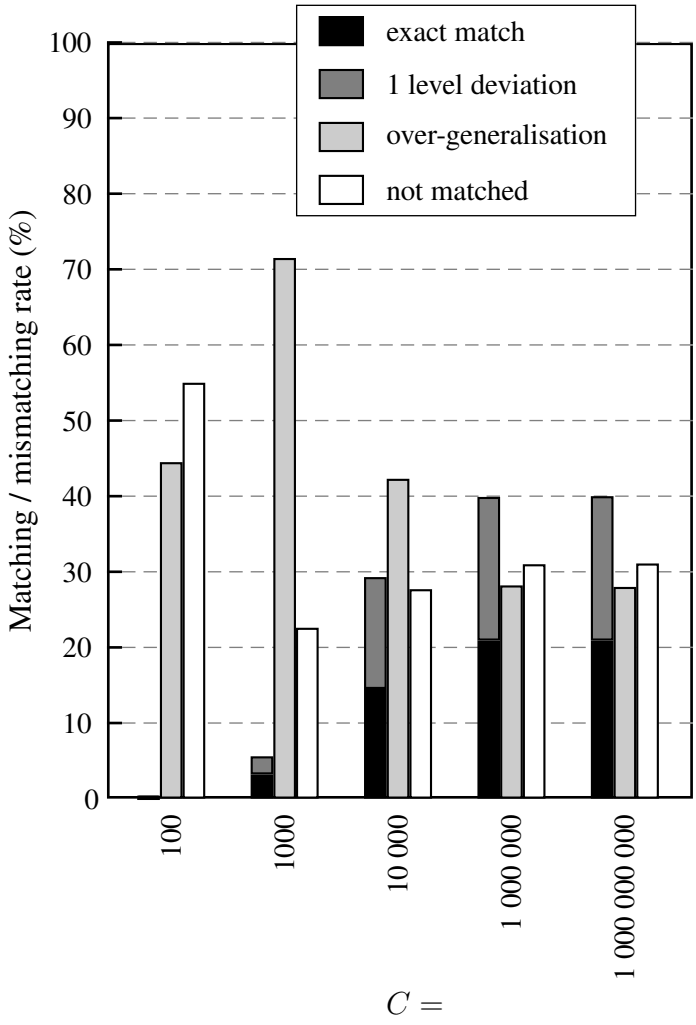


Figure 7.2: Gold standard matching rates for AGENT applying different values of C without WSD

role type	AGENT			
C	100			
gold standard noun concepts which are...	number (percentage)			
	with WSD		without WSD	
exactly matched	0	(0.0%)	1	(0.1%)
matched by 1 level hyponym	0	(0.0%)	0	(0.0%)
matched by 1 level hyperonym	3	(0.5%)	3	(0.3%)
matched by $\geq$ 2 level hyponym	0	(0.0%)	0	(0.0%)
matched by $\geq$ 2 level hyperonym	399	(64.8%)	382	(44.5%)
not matched	214	(34.7%)	472	(55.0%)
$\sum$	616		858	
acquired noun concepts per verb	4.4		2.5	

Table 7.2: Basic performance for the AGENT role, $C = 100$

role type	AGENT	
C	1000	
gold standard noun concepts which are...	number (percentage) with WSD	without WSD
exactly matched	34 (4.6%)	35 (3.2%)
matched by 1 level hyponym	19 (2.6%)	14 (1.3%)
matched by 1 level hyperonym	28 (3.8%)	12 (1.1%)
matched by $\geq$ 2 level hyponym	7 (0.9%)	2 (0.2%)
matched by $\geq$ 2 level hyperonym	449 (60.8%)	771 (71.5%)
not matched	201 (27.2%)	244 (22.6%)
$\sum$	738	1078
acquired noun concepts per verb	37.7	33.8

Table 7.3: Basic performance for the AGENT role, $C = 1000$

role type	AGENT	
C	10 000	
gold standard noun concepts which are...	number (percentage) with WSD	without WSD
exactly matched	96 (13.0%)	161 (14.5%)
matched by 1 level hyponym	48 (6.5%)	77 (6.9%)
matched by 1 level hyperonym	58 (7.9%)	87 (7.9%)
matched by $\geq$ 2 level hyponym	10 (1.4%)	7 (0.6%)
matched by $\geq$ 2 level hyperonym	257 (34.8%)	469 (42.3%)
not matched	269 (36.4%)	307 (27.7%)
$\sum$	738	1108
acquired noun concepts per verb	106.9	176.2

Table 7.4: Basic performance for the AGENT role, $C = 10\ 000$

role type	AGENT	
C	1 000 000	
gold standard noun concepts which are...	number (percentage) with WSD	without WSD
exactly matched	108 (14.6%)	232 (20.9%)
matched by 1 level hyponym	57 (7.7%)	102 (9.2%)
matched by 1 level hyperonym	70 (9.5%)	109 (9.8%)
matched by $\geq$ 2 level hyponym	11 (1.5%)	9 (0.8%)
matched by $\geq$ 2 level hyperonym	203 (27.5%)	312 (28.2%)
not matched	289 (39.2%)	344 (31.0%)
$\sum$	738	1108
acquired noun concepts per verb	138.1	249.1

Table 7.5: Basic performance for the AGENT role, $C = 1\ 000\ 000$

role type	AGENT			
C	1 000 000 000			
gold standard noun concepts which are...	number (percentage)			
	with WSD		without WSD	
exactly matched	108	(14.6%)	232	(20.9%)
matched by 1 level hyponym	57	(7.7%)	102	(9.2%)
matched by 1 level hyperonym	70	(9.5%)	110	(9.9%)
matched by ≥ 2 level hyponym	11	(1.5%)	9	(0.8%)
matched by ≥ 2 level hyperonym	202	(27.4%)	310	(28.0%)
not matched	290	(39.3%)	345	(31.1%)
$\sum$	738		1108	
acquired noun concepts per verb	138.5		250.6	

Table 7.6: Basic performance for the AGENT role, $C = 1\ 000\ 000\ 000$

28% of the matching tree cut concepts are too general. Furthermore, the results hardly differ between $C = 1\ 000\ 000$ and $C = 1\ 000\ 000\ 000$, although the latter value exceeds the former one by a factor of 1000. This means that the lowest possible generalisation level (where the synsets on the cuts immediately correspond to senses of the nouns in the data) is reached in most cases with $C = 1\ 000\ 000$, and specificity cannot be increased any more by raising the constant value. To understand this apparently contradictory behaviour, let us look at some examples. One gold standard relation is

(7.4) <rescue> INVOLVED_AGENT <liberator>

Figure 7.3 shows a part of the tree cut model acquired for <rescue> ($C = 1\ 000\ 000$, without WSD). As one can see, the displayed concepts are very specific; they are at the same hierarchical level as <liberator> (2 hyponym steps from <person>) or even below. Hence, <liberator> is not missing on the cut due to inappropriate generalisation. The reason is much simpler: "liberator" does not occur as Agent of "rescue" in the data. We have already encountered examples of very specific noun concepts in the gold standard. The results presented here suggests two conclusions: firstly, such specific noun concepts pose a serious data sparseness problem—the danger of their absence in the data is quite high; secondly, a large portion of AGENT relations in the gold standard make use of such concepts. I will examine the latter point in more detail in section 7.8.

The considerations in the previous paragraph explain the high percentage of concepts not matched at all. However, they still do not account for the finding that even cuts at a low abstraction level match the gold standard to a large extent by too general concepts. Again, to elucidate this I start with an example. The gold standard contains the following AGENT relations for <travel>:

(7.5)

a. <travel> INVOLVED_AGENT <traveler#traveller>

b. <travel> INVOLVED_AGENT <voyager>

concept	pref. value
...	...
<headmaster#schoolmaster#master>	27.66
<appointee#appointment>	16.00
<destroyer#ruiner>	310.85
<owner#proprietor>	5.53
<owner#possessor>	3.99
<cipher#cypher#nobody#nonentity>	16.10
<farmer#husbandman#granger>	9.04
<champion#fighter#hero#paladin>	15.67
<custodian#keeper#steward>	8.08
<fireman#fire_fighter#fire-fighter#fire_eater>	485.60
<prison_guard#jailer#gaoler#keeper#screw#turnkey>	35.25
<policeman#police_officer#officer>	14.76
<combatant#battler#belligerent#fighter#scrapper>	12.16
...	...

Figure 7.3: Part of the tree cut model for <rescue>

c. <travel> INVOLVED_AGENT <wanderer#roamer#rover>

d. <travel> INVOLVED_AGENT <salesman>

Figure 7.4 shows a part of the tree cut model acquired for <travel> ($C = 1\ 000\ 000$, without WSD). This part contains several virtual leaves. A virtual leaf is a (virtual) copy of an inner node in the hierarchy which is treated as a hyponym of that inner node (cf. sections 4.1.1.1 and 5.3). This is necessary since the tree cut approach requires that each word sense is represented by a leaf. A virtual leaf represents a sense of the words in the synset of the respective concept. Thus, a virtual leaf captures the "rest" of word senses which are subsumed by the corresponding inner node but not encoded by one of its hyponyms. In the figure, virtual leaves are indicated by the prefix 'REST::'. For example, regarding the concepts in figure 7.4, <traveler#traveller> is a hyperonym of, among others, <passenger#rider>, <guest#invitee>, and <migrant>. The latter concept, in turn, is a hyperonym of <immigrant>. The virtual leaves <REST::traveler#traveller> and <REST::migrant> represent a sense of "traveler" and "traveller", or "migrant", respectively. The fact that e.g. <REST::migrant> has a high preference value means that "migrant" occurs as the Agent of "travel" with a probability that is significantly higher than its overall occurrence probability in the data. Hence, different levels of abstraction can be found in a tree cut model if these different levels occur in the data to a significant extent.

Comparing the gold standard relations and the acquired tree cut model for <travel> yields the following: relations (7.5 a.)–(7.5 c.) are matched by <REST::traveler#traveller> in the model—a. is exactly matched, b. and c. are approximately matched since <voyager> and <wanderer> are immediate hyponyms of <traveler>. Relation (7.5 d.) is matched by <REST::worker> in the model. <worker> subsumes <salesman>, being located three levels higher in the hierarchy. Hence, this match falls under the category "matched by ≥ 2 level hyperonym". This example illustrates that tree cut models might encompass various levels of generalisation introduced by virtual leaves. This explains the

concept	pref. value
...	...
<immigrant>	28.16
<REST::migrant>	48.94
<passenger#rider>	11.80
<guest#invitee>	5.77
<REST::traveler#traveller>	10.03
<body_servant>	2.29
<maid#maidservant#housemaid#amah	10.13
<valet#valet_de_chambre#gentleman#gentleman's_gentleman#man	3.95
<craftsman#artisan#journeyman#artificer>	1.36
<REST::worker>	3.71
...	...

Figure 7.4: Part of the tree cut model for <travel>

surprising effect that even very specific cuts match gold standard concepts at a too general level. The evaluation results indicate that this is a common case. Very specific concepts, which do not occur in the data (indeed, "salesman" does not occur as a complement of "travel" either), can often "at least" be captured by general virtual leaves.

The overall relevance of virtual leaves, their status, and their treatment within the evaluation experiments deserve further discussion. For example, I treat a match by a virtual leaf as a match by the corresponding inner node. This decision is justifiable, but not trivial. Section 7.7 will address virtual leaves in more detail.

As expected, the average number of acquired concepts per verb increase with C (visualised by the bar diagram in figure 7.5). Again, for $C > 1\ 000\ 000$, no noteworthy difference can be observed. If the cuts are already at the lowest possible abstraction level, the number of concepts has reached its maximum. In general, disambiguation reduces both accuracy and the number of concepts. As explained in section 7.2 above, this is the expected behaviour. Only for low values of C (100 and 1000), the reverse holds. It seems that in these cases, the accumulation of erroneous senses leads to inadequate results (cf. footnote 1 on page 213). For higher values of C, the reduction of concepts achieved by the WSD step is much more substantial than the accompanying loss of accuracy. For $C = 10\ 000$, WSD reduces the average number of concepts from 176.2 to 106.9, which means a decrease of about 39%. The corresponding decrease of accuracy is moderate (from 29.3% to 27.4% approximate hits). For $C = 1\ 000\ 000$, the drop of accuracy caused by WSD is more significant (39.9% vs. 31.8% approximate hits), but the reduction of concepts (about 45%) is even higher. As discussed already, this is an indirect indicator of the performance of the employed disambiguation strategy: most of the concepts that are dropped by disambiguation are—as it should be—erroneous senses.

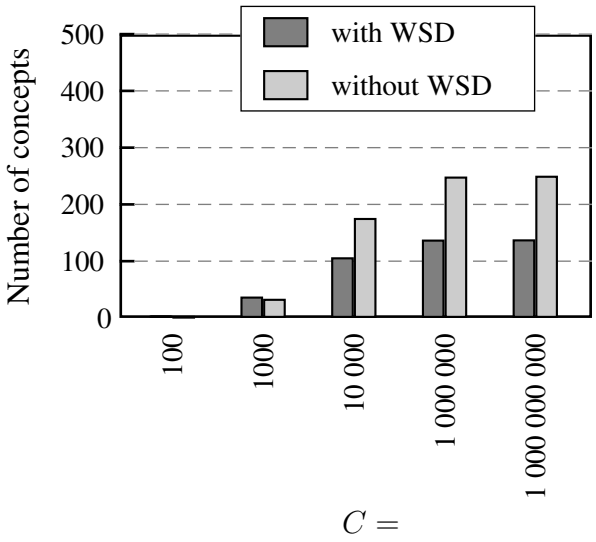


Figure 7.5: Average number of acquired concepts per verb for AGENT applying different values of C with and without WSD

7.3.2 PATIENT

Tables 7.7 to 7.11 show the results for the PATIENT role. Figures 7.6 and 7.7 provide corresponding bar diagrams of matching / mismatching rates. Compared to the results for the AGENT role, it is most striking that the accuracy results are much better: the maximal percentage of approximate matches is 69.1%, achieved with $C = 1\ 000\ 000$ (or higher) and without disambiguation. This is almost twice as much as the corresponding value for AGENT (39.9%). On the other hand, the amount of over-general matches and complete mismatches is rather low (13.4% and 13.0%, respectively, given the parameters just mentioned). As we will see later, the reason for this discrepancy is a heterogeneity in the gold standard across different role types: for PATIENT, a larger amount of encoded noun concepts is at a relatively general level, while such concepts occur to a lower percentage within AGENT relations. For example, although almost 90% of AGENT relations in the gold standard involve noun concepts subsumed by <person> (cf. section 6.5), only 7 AGENT relations contain this concept itself. In contrast, 29 PATIENT relations in the gold standard contain <person>, although only about 10% of them include noun concepts denoting some kind of person. This implies that very specific concepts, which pose the problems described in the previous section, play a minor role for PATIENT than they do for AGENT. I will examine the correlation between the generalisation level in the gold standard and the performance of the acquisition algorithm for the different role types in section 7.8.

Another major difference between the results for AGENT and PATIENT is that with the latter role type the average number of acquired concepts (visualised in figure 7.8) is much higher, namely up to 782.2 (with WSD) or 1743.9 (without WSD), as opposed to 138.5 / 250.6 for AGENT. This is an expected finding, since the semantic range of PATIENT is much more diverse than the one for AGENT so that no semantic filter for the PATIENT role can be employed (cf. section 6.5).

Apart from these differences, it turns out that in principle, most of the tendencies noted for the AGENT role can be observed here as well. Accuracy improves with growing C, as does the average number

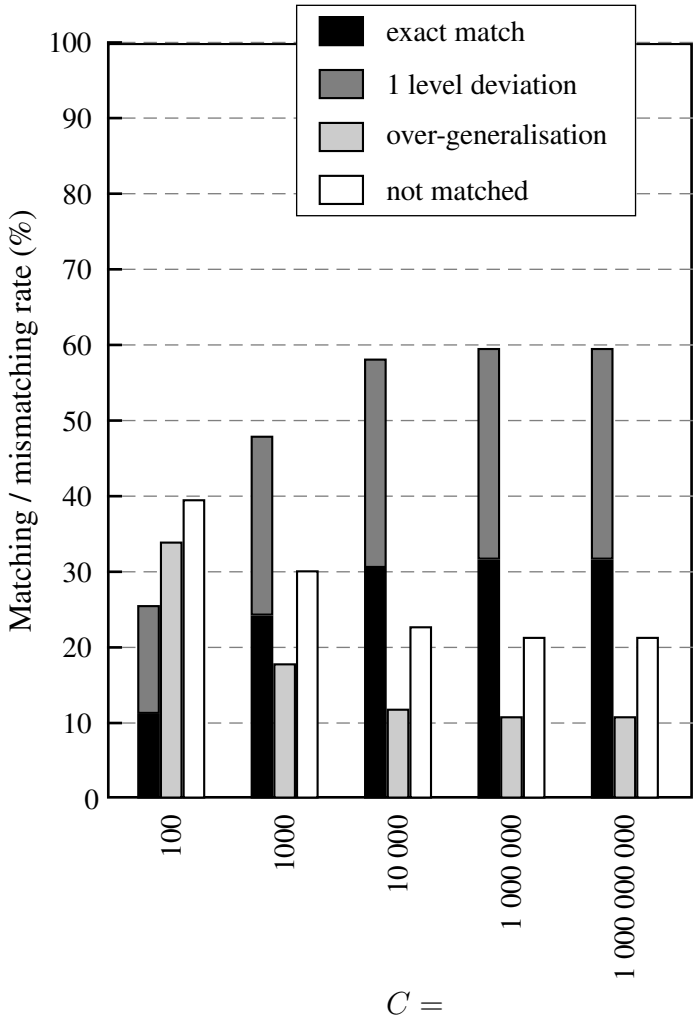


Figure 7.6: Gold standard matching rates for PATIENT applying different values of C with WSD

role type	PATIENT			
C	100			
gold standard noun concepts which are...	number (percentage)			
	with WSD		without WSD	
exactly matched	32	(11.2%)	39	(8.4%)
matched by 1 level hyponym	14	(4.9%)	6	(1.3%)
matched by 1 level hyperonym	27	(9.5%)	47	(10.2%)
matched by $\geq$ 2 level hyponym	2	(0.7%)	1	(0.2%)
matched by $\geq$ 2 level hyperonym	97	(34.0%)	193	(41.8%)
not matched	113	(39.6%)	176	(38.1%)
$\sum$	285		462	
acquired noun concepts per verb	29.8		21.7	

Table 7.7: Basic performance for the PATIENT role, $C = 100$

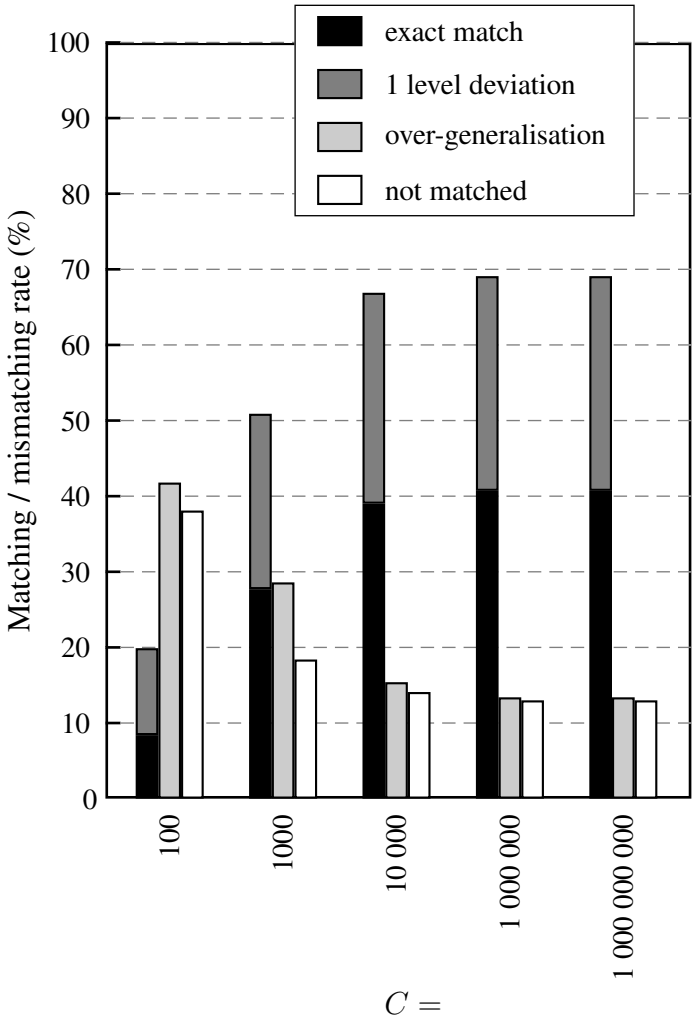


Figure 7.7: Gold standard matching rates for PATIENT applying different values of C without WSD

role type	PATIENT	
C	1000	
gold standard noun concepts which are...	number (percentage) with WSD	without WSD
exactly matched	69 (24.2%)	128 (27.7%)
matched by 1 level hyponym	30 (10.5%)	43 (9.3%)
matched by 1 level hyperonym	38 (13.3%)	64 (13.9%)
matched by $\geq$ 2 level hyponym	11 (3.9%)	10 (2.2%)
matched by $\geq$ 2 level hyperonym	51 (17.9%)	132 (28.6%)
not matched	86 (30.2%)	85 (18.4%)
$\sum$	285	462
acquired noun concepts per verb	242.5	339.3

Table 7.8: Basic performance for the PATIENT role, $C = 1000$

role type	PATIENT	
C	10 000	
gold standard noun concepts which are...	number (percentage) with WSD	without WSD
exactly matched	87 (30.5%)	180 (39.0%)
matched by 1 level hyponym	54 (18.9%)	81 (17.5%)
matched by 1 level hyperonym	25 (8.8%)	48 (10.4%)
matched by ≥ 2 level hyponym	20 (7.0%)	17 (3.7%)
matched by ≥ 2 level hyperonym	34 (11.9%)	71 (15.4%)
not matched	65 (22.8%)	65 (14.1%)
$\sum$	285	462
acquired noun concepts per verb	611.1	1372.0

Table 7.9: Basic performance for the PATIENT role, $C = 10\ 000$

role type	PATIENT	
C	1 000 000	
gold standard noun concepts which are...	number (percentage) with WSD	without WSD
exactly matched	90 (31.6%)	188 (40.7%)
matched by 1 level hyponym	54 (18.9%)	89 (19.3%)
matched by 1 level hyperonym	26 (9.1%)	42 (9.1%)
matched by ≥ 2 level hyponym	23 (8.1%)	21 (4.5%)
matched by ≥ 2 level hyperonym	31 (10.9%)	62 (13.4%)
not matched	61 (21.4%)	60 (13.0%)
$\sum$	285	462
acquired noun concepts per verb	772.8	1739.4

Table 7.10: Basic performance for the PATIENT role, $C = 1\ 000\ 000$

role type	PATIENT	
C	1 000 000 000	
gold standard noun concepts which are...	number (percentage) with WSD	without WSD
exactly matched	90 (31.6%)	188 (40.7%)
matched by 1 level hyponym	54 (18.9%)	88 (19.0%)
matched by 1 level hyperonym	26 (9.1%)	43 (9.3%)
matched by ≥ 2 level hyponym	23 (8.1%)	21 (4.5%)
matched by ≥ 2 level hyperonym	31 (10.9%)	62 (13.4%)
not matched	61 (21.4%)	60 (13.0%)
$\sum$	285	462
acquired noun concepts per verb	782.2	1743.9

Table 7.11: Basic performance for the PATIENT role, $C = 1\ 000\ 000\ 000$

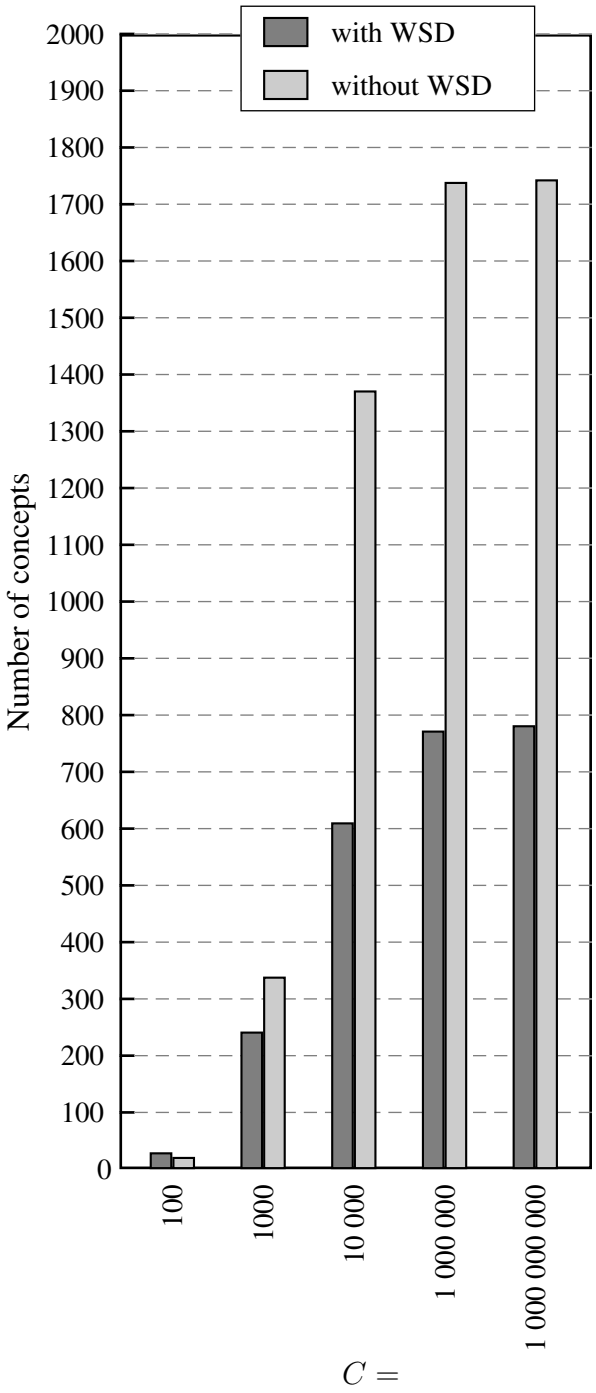


Figure 7.8: Average number of acquired concepts per verb for PATIENT applying different values of C with and without WSD

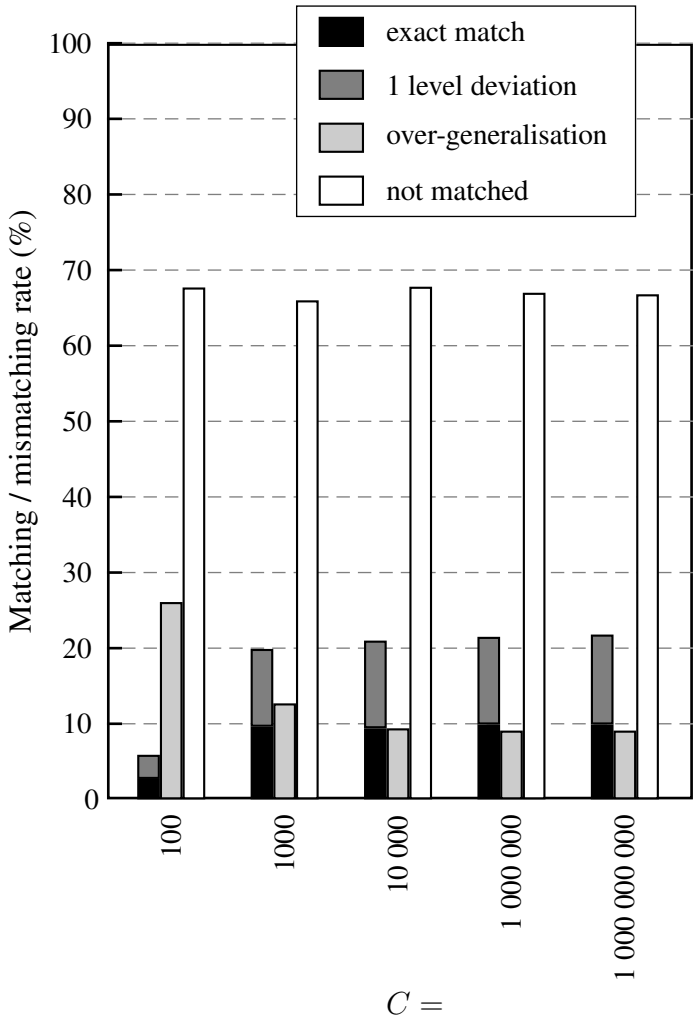


Figure 7.9: Gold standard matching rates for INSTRUMENT applying different values of C with WSD

of noun concepts per verb; however, no significant changes take place for $C > 1\ 000\ 000$, indicating that the lowest abstraction level has been reached. Except for $C = 100$, disambiguation results in a substantial reduction of the number of concepts (more than a half for $C \geq 10\ 000$), accompanied by a much less dramatic decrease of accuracy (e.g. from 69.1% to 59.6% for $C = 1\ 000\ 000$).

7.3.3 INSTRUMENT

Tables 7.12 to 7.16 show the results for the INSTRUMENT role; figures 7.9 to 7.11 provide the corresponding bar diagrams. Major tendencies that could be observed for the role types in the previous sections apply here as well: increasing C improves accuracy and increases the number of noun concepts per verb up to a certain point (1 000 000) beyond which no noteworthy changes take place; WSD decreases accuracy to some extent (except for $C \leq 1000$) while reducing the number of noun concepts substantially (except for $C = 100$). The most striking observation here is that accuracy is

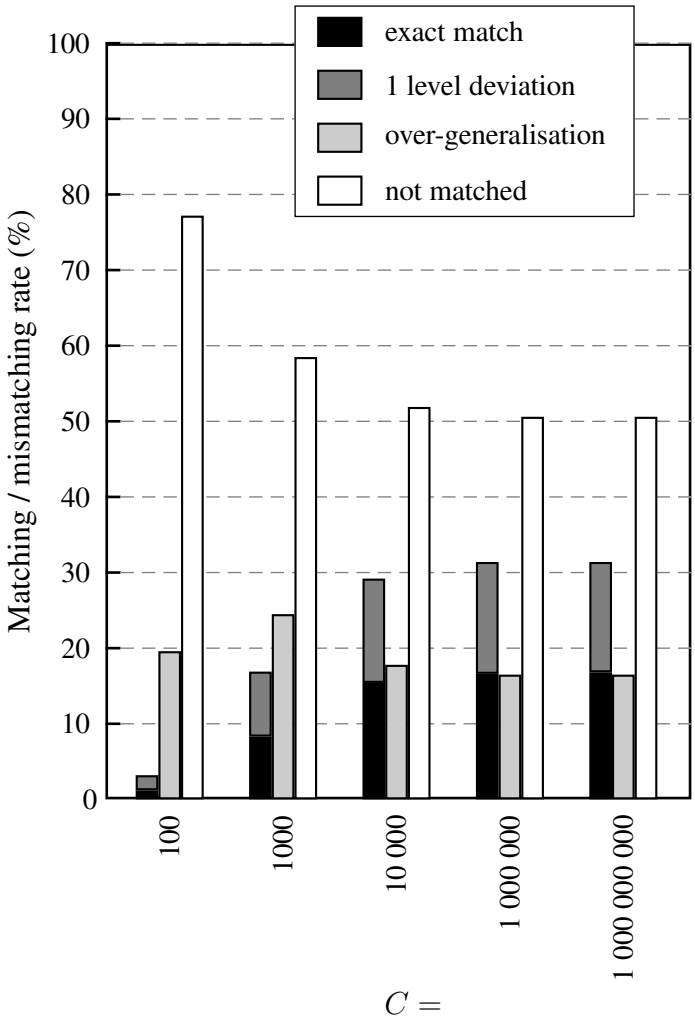


Figure 7.10: Gold standard matching rates for INSTRUMENT applying different values of C without WSD

even worse than for the AGENT role: the highest achieved rate of approximate hits is 31.4%, opposed to 40% with AGENTs. More seriously, the percentage of entirely unmatched concepts is significantly higher (50.6% or more; the corresponding rate for AGENT varies between about 20% and 30% for $C \geq 1000$ without WSD).

A manual inspection of a random sample of the unmatched gold standard relations makes this drop of accuracy become plausible. Obviously, a major portion of INSTRUMENT relations encode incorporations which do not coincide with derivational patterns so that the filter mechanism sketched in section 7.1 is not able to eliminate them. Some examples are listed in (7.6) to (7.12):

(7.6) <step#take_a_step> INVOLVED_INSTRUMENT <leg>

(7.7) <jump#leap#bound#spring> INVOLVED_INSTRUMENT <oil>

(7.8) <embark> INVOLVED_INSTRUMENT <landing_stage>

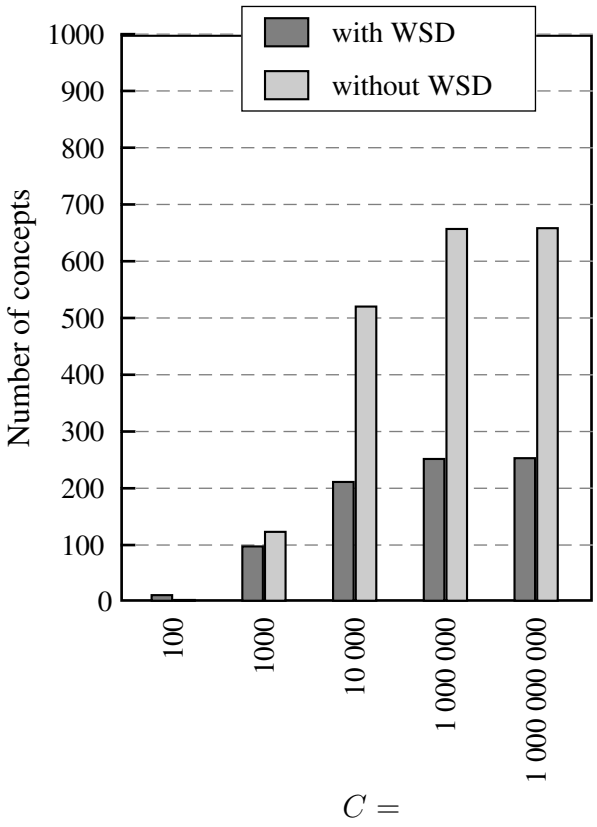


Figure 7.11: Average number of acquired concepts per verb for INSTRUMENT applying different values of C with and without WSD

role type	INSTRUMENT			
C	100			
gold standard noun concepts which are...	number (percentage)			
	with WSD		without WSD	
exactly matched	10	(2.7%)	6	(1.2%)
matched by 1 level hyponym	0	(0.0%)	1	(0.2%)
matched by 1 level hyperonym	12	(3.2%)	9	(1.8%)
matched by ≥ 2 level hyponym	1	(0.3%)	0	(0.0%)
matched by ≥ 2 level hyperonym	97	(26.1%)	101	(19.6%)
not matched	252	(67.7%)	397	(77.2%)
$\sum$	372		514	
acquired noun concepts per verb	13.6		5.1	

Table 7.12: Basic performance for the INSTRUMENT role, $C = 100$

role type	INSTRUMENT			
C	1000			
gold standard noun concepts which are...	number (percentage) with WSD		without WSD	
exactly matched	40	(9.6%)	59	(8.3%)
matched by 1 level hyponym	11	(2.6%)	19	(2.7%)
matched by 1 level hyperonym	32	(7.7%)	42	(5.9%)
matched by $\geq$ 2 level hyponym	6	(1.4%)	0	(0.0%)
matched by $\geq$ 2 level hyperonym	53	(12.7%)	174	(24.5%)
not matched	276	(66.0%)	415	(58.5%)
$\sum$	418		709	
acquired noun concepts per verb	99.4		125.1	

Table 7.13: Basic performance for the INSTRUMENT role, $C = 1000$

role type	INSTRUMENT			
C	10 000			
gold standard noun concepts which are...	number (percentage) with WSD		without WSD	
exactly matched	37	(9.4%)	109	(15.4%)
matched by 1 level hyponym	23	(5.8%)	52	(7.3%)
matched by 1 level hyperonym	23	(5.8%)	46	(6.5%)
matched by $\geq$ 2 level hyponym	7	(1.8%)	8	(1.1%)
matched by $\geq$ 2 level hyperonym	37	(9.4%)	126	(17.8%)
not matched	267	(67.8%)	368	(51.9%)
$\sum$	394		709	
acquired noun concepts per verb	213.0		522.0	

Table 7.14: Basic performance for the INSTRUMENT role, $C = 10\ 000$

role type	INSTRUMENT			
C	1 000 000			
gold standard noun concepts which are...	number (percentage) with WSD		without WSD	
exactly matched	39	(9.9%)	118	(16.6%)
matched by 1 level hyponym	25	(6.3%)	58	(8.2%)
matched by 1 level hyperonym	21	(5.3%)	47	(6.6%)
matched by $\geq$ 2 level hyponym	9	(2.3%)	10	(1.4%)
matched by $\geq$ 2 level hyperonym	36	(9.1%)	117	(16.5%)
not matched	264	(67.0%)	359	(50.6%)
$\sum$	394		709	
acquired noun concepts per verb	253.5		658.7	

Table 7.15: Basic performance for the INSTRUMENT role, $C = 1\ 000\ 000$

role type	INSTRUMENT			
C	1 000 000 000			
gold standard noun concepts which are...	number (percentage)			
	with WSD		without WSD	
exactly matched	39	(9.9%)	119	(16.8%)
matched by 1 level hyponym	26	(6.6%)	57	(8.0%)
matched by 1 level hyperonym	21	(5.3%)	47	(6.6%)
matched by ≥ 2 level hyponym	9	(2.3%)	10	(1.4%)
matched by ≥ 2 level hyperonym	36	(9.1%)	117	(16.5%)
not matched	263	(66.8%)	359	(50.6%)
$\sum$	394		709	
acquired noun concepts per verb	254.9		660.1	

Table 7.16: Basic performance for the INSTRUMENT role, $C = 1\ 000\ 000\ 000$

(7.9) <pour#cause_to_run> INVOLVED_INSTRUMENT <pitcher#ewer>

(7.10) <smell#inhale_odor_of> INVOLVED_INSTRUMENT <nose#olfactory_organ>

(7.11) <pay#make_a_payment> INVOLVED_INSTRUMENT
<medium_of_exchange#monetary_system>

(7.12) <offer#proffer#present_for_acceptance> INVOLVED_INSTRUMENT <word>

The nouns in these relations (some of which appear somewhat unconventional to me) are hardly, if at all, expressed as instrumental complements of the corresponding verbs. Thus, they cannot be acquired from corpus data.

7.3.4 LOCATION

Tables 7.17 to 7.21 show results for the LOCATION role. I list these results for the sake of completeness, but will not draw substantial conclusions from them. The reason for this becomes clear if one looks at the number of gold standard relations which are included for evaluation (the row indicated by "$\sum$"): From the 173 LOCATION relations in the gold standard which remain after filtering derivational patterns (cf. table 7.1 on page 211), only 19 (with WSD) or, respectively, 27 (without WSD) relations pertain to verb concepts which appear in the data with sufficient frequency to be taken into account.[2] This is an unexpected finding. Unfortunately, this is not enough to be suitable as a basis for evaluation. A difference of matching or not matching one single relation amounts to a difference in accuracy of about 5% with disambiguation and 4% without disambiguation. Therefore, the results

[2]I would like to emphasise that the frequency criterion I employ (verb concepts must have a frequency of at least 50) is not primarily responsible for this discrepancy. The 173 LOCATION relations pertain to 116 different verb concepts altogether. From these, only 42 concepts without WSD and 25 concepts with WSD occur in the training data (or, more precisely, are extracted from them) at all.

role type	LOCATION	
C	100	
gold standard noun concepts which are...	number (percentage) with WSD	without WSD
exactly matched	4 (21.1%)	2 (8.0%)
matched by 1 level hyponym	0 (0.0%)	0 (0.0%)
matched by 1 level hyperonym	2 (10.5%)	3 (12.0%)
matched by ≥ 2 level hyponym	0 (0.0%)	0 (0.0%)
matched by ≥ 2 level hyperonym	10 (52.6%)	13 (52.0%)
not matched	3 (15.8%)	7 (28.0%)
$\sum$	19	25
acquired noun concepts per verb	16.2	10.5

Table 7.17: Basic performance for the LOCATION role, $C = 100$

would be rather unreliable, depending too much on idiosyncratic factors. Therefore, I will exclude the LOCATION type from further evaluation.

For the experiments described in the following sections, the value of C will be fixed to 1 000 000. For all role types, this value essentially yields the best accuracy rates.

7.4 The Impact of Semantic Filtering

In this section, I examine the impact of employing semantic filters. Recall that semantic filters are applied to the acquired tree cuts to narrow down their semantic range as adequate for the respective role type (cf. section 6.5). For each role type, the corresponding semantic filter demarcates a certain part of the WordNet hierarchy which represents the typical semantic range of that role. (For example, for the AGENT role, this range comprises all concepts subsumed by <life_form>, <causal_agent>, and <group#grouping>.) From a tree cut model acquired for a certain role, the corresponding filter selects the concepts which are within this range and discards the other concepts on the cut. To test the impact of such filters, I performed the evaluation experiments without applying semantic filters. Tables 7.22 and 7.23 show the results for AGENT and INSTRUMENT, respectively. (Recall that for PATIENT, no semantic filter is employed.) Figures 7.12 and 7.13 visualise the comparison of accuracy and average number of acquired concepts with and without filtering.

Comparing the AGENT results in table 7.22 with the corresponding results in section 7.3 (i.e. table 7.5 on page 216), one can state two facts: firstly, semantic filtering decreases accuracy to a fairly low extent. The rate of approximate matches is reduced from 32.8% to 31.8% with WSD and from 42.5% to 39.9% without WSD. This finding is expectable, since the filters are defined in a way that at least 90% of the gold standard relations of each role type pass the filter. Thus, the relations eliminated by the filter could only cover a small amount of gold standard relations so that their absence does not

role type	LOCATION	
C	1000	
gold standard noun concepts which are...	number (percentage) with WSD	without WSD
exactly matched	6 (31.6%)	5 (18.5%)
matched by 1 level hyponym	2 (10.5%)	2 (7.4%)
matched by 1 level hyperonym	1 (5.3%)	1 (3.7%)
matched by ≥ 2 level hyponym	0 (0.0%)	0 (0.0%)
matched by ≥ 2 level hyperonym	1 (5.3%)	13 (48.1%)
not matched	9 (47.4%)	6 (22.2%)
$\sum$	19	27
acquired noun concepts per verb	87.0	92.4

Table 7.18: Basic performance for the LOCATION role, $C = 1000$

role type	LOCATION	
C	10 000	
gold standard noun concepts which are...	number (percentage) with WSD	without WSD
exactly matched	8 (42.1%)	8 (29.6%)
matched by 1 level hyponym	3 (15.8%)	4 (14.8%)
matched by 1 level hyperonym	1 (5.3%)	5 (18.5%)
matched by ≥ 2 level hyponym	0 (0.0%)	0 (0.0%)
matched by ≥ 2 level hyperonym	0 (0.0%)	4 (14.8%)
not matched	7 (36.8%)	6 (22.2%)
$\sum$	19	27
acquired noun concepts per verb	172.5	486.7

Table 7.19: Basic performance for the LOCATION role, $C = 10\ 000$

role type	LOCATION	
C	1 000 000	
gold standard noun concepts which are...	number (percentage) with WSD	without WSD
exactly matched	8 (42.1%)	8 (29.6%)
matched by 1 level hyponym	3 (15.8%)	5 (18.5%)
matched by 1 level hyperonym	1 (5.3%)	4 (14.8%)
matched by ≥ 2 level hyponym	0 (0.0%)	0 (0.0%)
matched by ≥ 2 level hyperonym	0 (0.0%)	4 (14.8%)
not matched	7 (36.8%)	6 (22.2%)
$\sum$	19	27
acquired noun concepts per verb	193.4	606.9

Table 7.20: Basic performance for the LOCATION role, $C = 1\ 000\ 000$

role type	LOCATION	
C	1 000 000	
gold standard noun concepts which are...	number (percentage) with WSD	without WSD
exactly matched	8 (42.1%)	8 (29.6%)
matched by 1 level hyponym	3 (15.8%)	5 (18.5%)
matched by 1 level hyperonym	1 (5.3%)	4 (14.8%)
matched by ≥ 2 level hyponym	0 (0.0%)	0 (0.0%)
matched by ≥ 2 level hyperonym	0 (0.0%)	4 (14.8%)
not matched	7 (36.8%)	6 (22.2%)
$\sum$	19	27
acquired noun concepts per verb	194.5	608.4

Table 7.21: Basic performance for the LOCATION role, $C = 1\ 000\ 000\ 000$

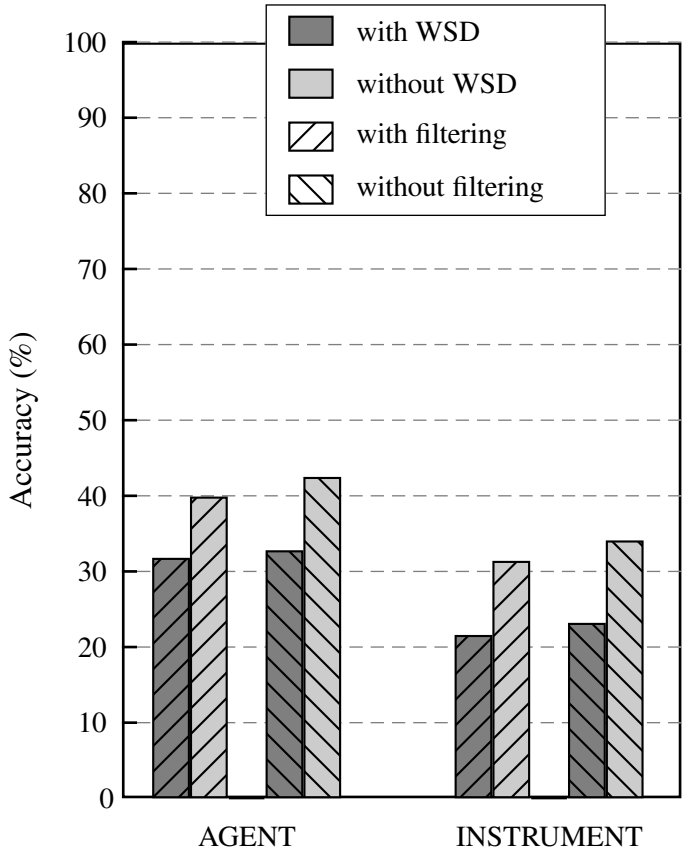


Figure 7.12: Accuracy (approximate match rates) for AGENT and INSTRUMENT with and without semantic filtering

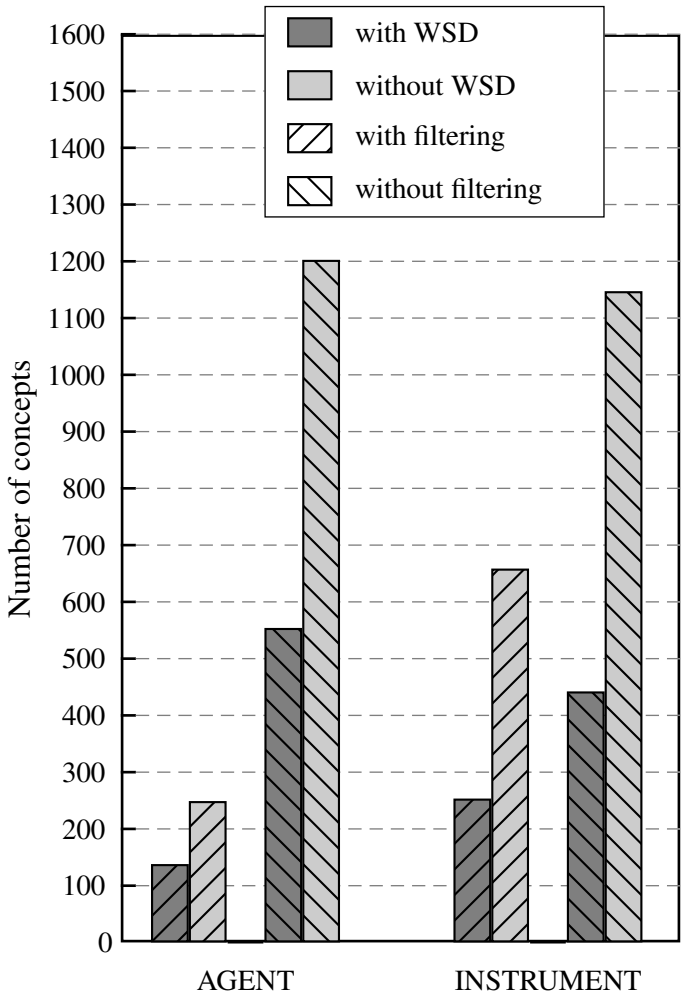


Figure 7.13: Average number of acquired concepts per verb for AGENT and INSTRUMENT with and without semantic filtering

role type	AGENT			
C	1 000 000			
gold standard noun concepts which are...	number (percentage)			
	with WSD		without WSD	
exactly matched	112	(15.2%)	251	(22.7%)
matched by 1 level hyponym	59	(8.0%)	108	(9.7%)
matched by 1 level hyperonym	71	(9.6%)	112	(10.1%)
matched by $\geq$ 2 level hyponym	13	(1.8%)	9	(0.8%)
matched by $\geq$ 2 level hyperonym	215	(29.1%)	330	(29.8%)
not matched	268	(36.3%)	298	(26.9%)
$\sum$	738		1108	
acquired noun concepts per verb	554.1		1202.6	

Table 7.22: Results for the AGENT role without semantic filtering

role type	INSTRUMENT			
C	1 000 000			
gold standard noun concepts which are...	number (percentage)			
	with WSD		without WSD	
exactly matched	42	(10.7%)	128	(18.1%)
matched by 1 level hyponym	27	(6.9%)	63	(8.9%)
matched by 1 level hyperonym	22	(5.6%)	50	(7.1%)
matched by $\geq$ 2 level hyponym	9	(2.3%)	12	(1.7%)
matched by $\geq$ 2 level hyperonym	44	(11.2%)	144	(20.3%)
not matched	250	(63.5%)	312	(44.0%)
$\sum$	394		709	
acquired noun concepts per verb	442.3		1147.4	

Table 7.23: Results for the INSTRUMENT role without semantic filtering

have a major effect on accuracy. On the other hand, semantic filtering leads to a dramatic reduction of the acquired noun concepts per verb: from 554.1 to 138.1 with WSD and from 1202.6 to 249.1 without WSD. This is a reduction of more than 75%.

For INSTRUMENT, the impact of semantic filtering is comparable: The comparison of table 7.23 with the corresponding table 7.15 on page 228 shows that applying the filter yields a small reduction of approximate matches form 23.2% to 21.6% with WSD and from 34.1% to 31.4% without WSD. Conversely, the average number of noun concepts decreases significantly, namely from 442.3 to 253.5 with WSD and from 1147.4 to 658.7 without WSD, a reduction of about 45%.[3]

Overall, one can state that the semantic filters work as intended. The significant reduction of acquired "candidate relations" is very useful within the semi-automatic acquisition scenario, since this immediately reduces the effort of manual inspection of these candidates. This outweighs the moderate loss of accuracy semantic filtering brings about.

7.5 The Impact of Argument Clustering

After testing the impact of semantic filtering, this section examines the effect that the approach of clustering arguments and determining their roles developed in chapter 6 (section 6.2 to section 6.4) has on the acquired results. Instead of employing this rather sophisticated clustering strategy to map syntactic arguments to thematic roles, one could apply simplistic heuristics like *Use all subjects to acquire AGENT relations* or, as in section 5.5, *Use all objects to acquire PATIENT relations*. In the following, I test how the results look like if these two heuristics are employed.

Comparing these two rules to the argument clustering approach is particularly interesting because the latter is supposed to cope with a common phenomenon which the simple rules do not take into account, namely the causative/inchoative alternation. As exhaustively discussed in the previous chapter, the linking approach should recognise that e.g. the subject in the intransitive sentence "The vase broke." (which corresponds to the object in the analogous transitive sentence "The man broke the vase.") realises a Patient so that the pair *(break, vase)* serves as input for learning PATIENT relations. In contrast, the simplifying rules just stated employ *(break, vase)* for acquiring AGENT relations. However, the semantic filter would eliminate the corresponding concept(s). Moreover, only if "vase" also occurred as object of "break" in the data, then the simple heuristics would also take this pair into account for learning PATIENT relations. However, if "vase" did not occur as object of "break", but nouns similar to "vase" did, then the clustering strategy still could recognise that this noun expresses a Patient though it is only found in subject position.

These considerations lead to the following expectations regarding the performance with clustering vs. with simple linking heuristics: For AGENT, there should be no significant differences; the heuristics include some data items which are inadequate, but the corresponding concepts are filtered out. For PATIENT, the clustering technique includes data items which are missed by the heuristics. Therefore, clustering should yield a higher coverage of the gold standard relations and maybe improve accuracy.

Tables 7.24 and 7.25 show the results obtained by using the above-mentioned heuristics. A comparison with the corresponding tables 7.5 on page 216 and 7.10 on page 223 yields the following: Overall,

[3]This reduction rate does not reach the corresponding rate for AGENT, because the semantic range determined for the AGENT role is much narrower than the range for the INSTRUMENT role, cf. section 6.5.

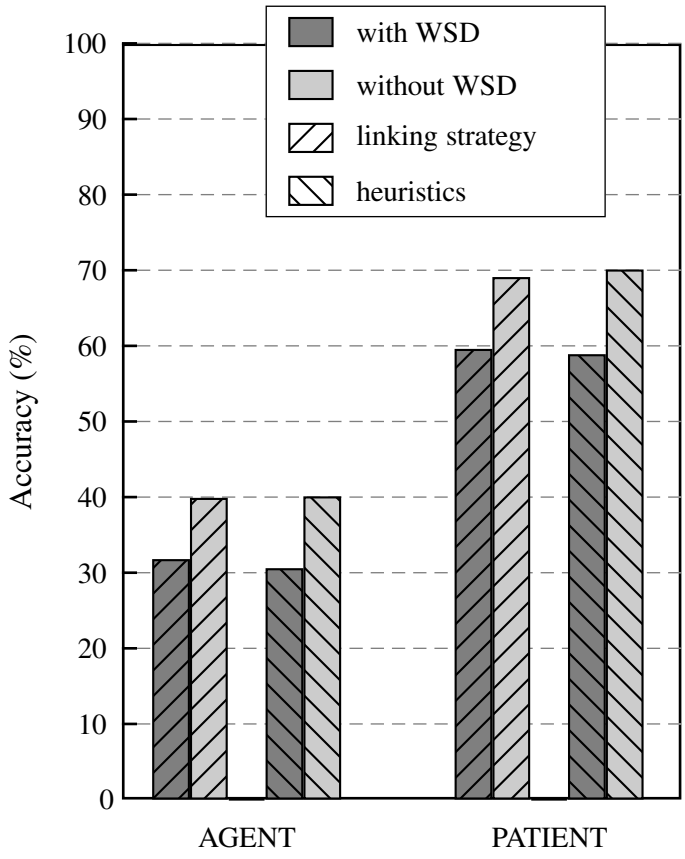


Figure 7.14: Accuracy (approximate match rates) for AGENT and PATIENT with the linking strategy and with simple heuristics

role type	AGENT	
C	1 000 000	
gold standard noun concepts which are...	number (percentage) with WSD	without WSD
exactly matched	108 (14.1%)	239 (20.8%)
matched by 1 level hyponym	57 (7.5%)	112 (9.8%)
matched by 1 level hyperonym	69 (9.0%)	109 (9.5%)
matched by $\geq$ 2 level hyponym	14 (1.8%)	10 (0.9%)
matched by $\geq$ 2 level hyperonym	218 (28.5%)	318 (27.7%)
not matched	298 (39.0%)	360 (31.4%)
$\sum$	764	1148
acquired noun concepts per verb	135.7	248.5

Table 7.24: Results for the AGENT role, drawn from all subjects

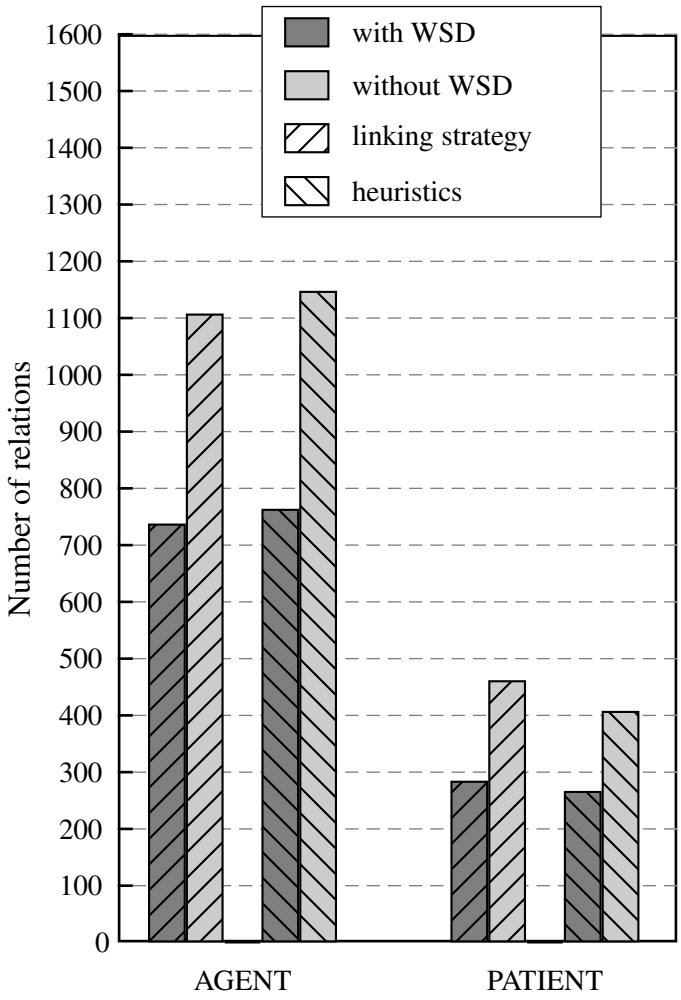


Figure 7.15: Coverage of the gold standard (number of relations taken into account for evaluation) for AGENT and PATIENT with the linking strategy and with simple heuristics

role type	PATIENT	
C	1 000 000	
gold standard noun concepts which are...	number (percentage) with WSD	without WSD
exactly matched	80 (30.0%)	166 (40.7%)
matched by 1 level hyponym	49 (18.4%)	77 (18.9%)
matched by 1 level hyperonym	28 (10.5%)	43 (10.5%)
matched by $\geq$ 2 level hyponym	18 (6.7%)	17 (4.2%)
matched by $\geq$ 2 level hyperonym	33 (12.4%)	52 (12.7%)
not matched	59 (22.1%)	53 (13.0%)
$\sum$	267	408
acquired noun concepts per verb	669.4	1530.0

Table 7.25: Results for the PATIENT role, drawn from objects only

the accuracy rates are fairly similar (cf. figure 7.14). Thus, the clustering strategy neither improves nor impairs performance significantly in this respect. However, there are remarkable changes concerning the coverage of the gold standard (row in the tables labelled with "$\sum$"), i.e. the number of gold standard relations which can be taken into account for evaluation (cf. figure 7.15). Recall that, as motivated in section 7.2, the evaluation procedure considers only those gold standard relations that contain a verb concept for which the learning algorithm effectively does acquire any relations of the inspected role type. Without WSD, the number of considered gold standard relations increases from 408 to 462. As explained above, this behaviour is expected since more data items are taken into account. Along with that it is not surprising that the average number of acquired noun concepts is significantly higher with clustering (772.8 / 1739.4 vs. 669.4 / 1530). Regarding AGENT, it is striking that clustering decreases the coverage of gold standard relations (e.g. without WSD from 1148 to 1108). Obviously, some data items which really represent Agents are erroneously misclassified by the clustering approach. However, taking into account that there are twice as many AGENT relations as PATIENT relations in the gold standard (cf. table 7.1 on page 211), the coverage increase of PATIENT relations outweighs the coverage decrease of AGENT relations.

7.6 The Impact of the LSC Model

One important component of the approach of learning role relations proposed in this thesis is the latent semantic class (LSC) model (cf. section 3.3). This model is applied twice; once for word sense disambiguation (cf. section 5.2.2) and once within the linking approach, to cluster similar arguments of a verb (cf. section 6.2). For training an LSC model, one has to fix two variables in advance: the number of latent semantic classes and the number of training iterations. The LSC model I used so far comprises 35 classes and was trained with 400 iterations. One might expect that changing these variables affects the performance of the WSD and/or the linking component. For example, a higher number of classes and iterations yields a more fine-grained model that might increase the performance of WSD and/or argument clustering (because this model exhibits a higher discriminative power) or decrease this performance (because the larger amount of parameters to be estimated causes a sparse-

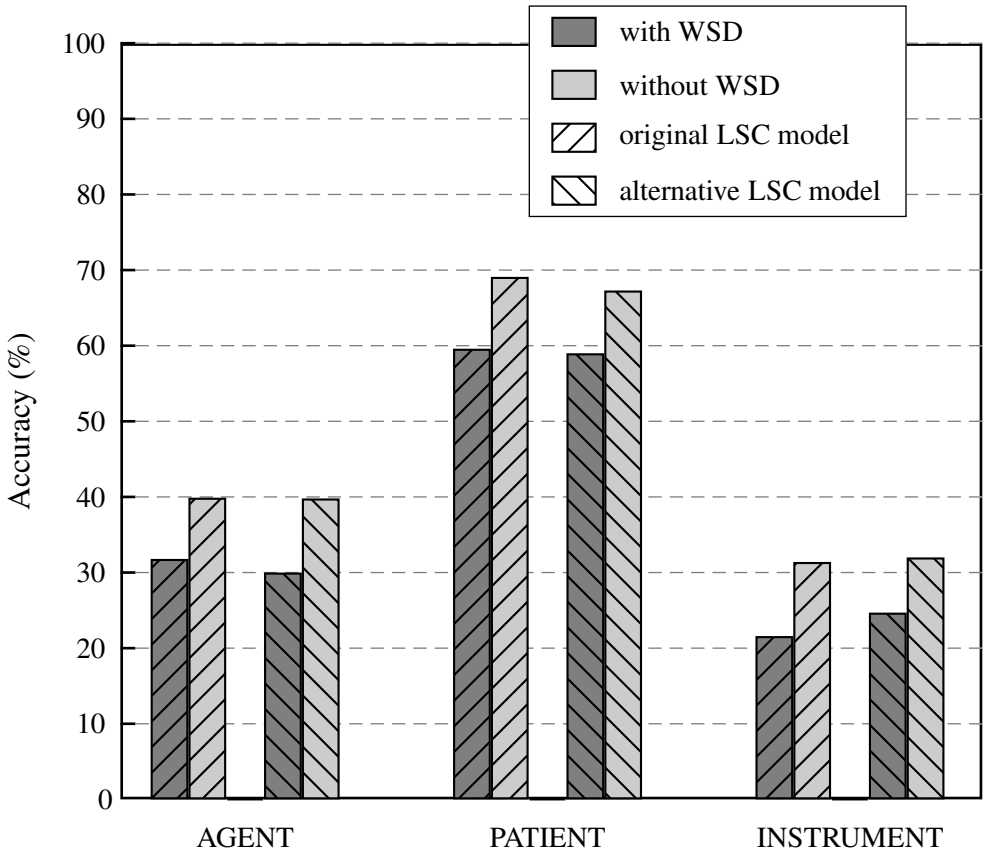


Figure 7.16: Accuracy (approximate match rates) for AGENT, PATIENT, and INSTRUMENT with the original and an alternative LSC model

data problem). To test whether the setting of these variables has an effect on the performance of the acquisition approach, I did experiments which employ an alternative LSC model with different settings, namely 70 classes and 800 iterations. Tables 7.26 to 7.28 show the results for AGENT, PATIENT, and INSTRUMENT, respectively. The corresponding results with the "original" LSC model are listed in table 7.5 on page 216, table 7.10 on page 223, and table 7.15 on page 228. Figures 7.16 and 7.17 provide bar diagrams which allow an immediate comparison.

Overall, the results obtained with the two LSC models are comparable. The patterns of accuracy rates are very similar and the respective average numbers of noun concepts are very close. For AGENT and PATIENT, the originally employed model yields slightly better accuracy rates, while for INSTRUMENT, the results are slightly better using the alternative model. Likewise, the impact of changing the LSC model on the average number of concepts is nonuniform. Hence, no general tendency can be observed which favours one model over the other. This experiment provides no evidence for the hypothesis that the selection of the LSC model parameters plays a major role for our task. However, this is a preliminary finding. To ultimately disprove—or prove—a general relationship between the employed LSC model and performance, a systematic empirical investigation employing a larger num-

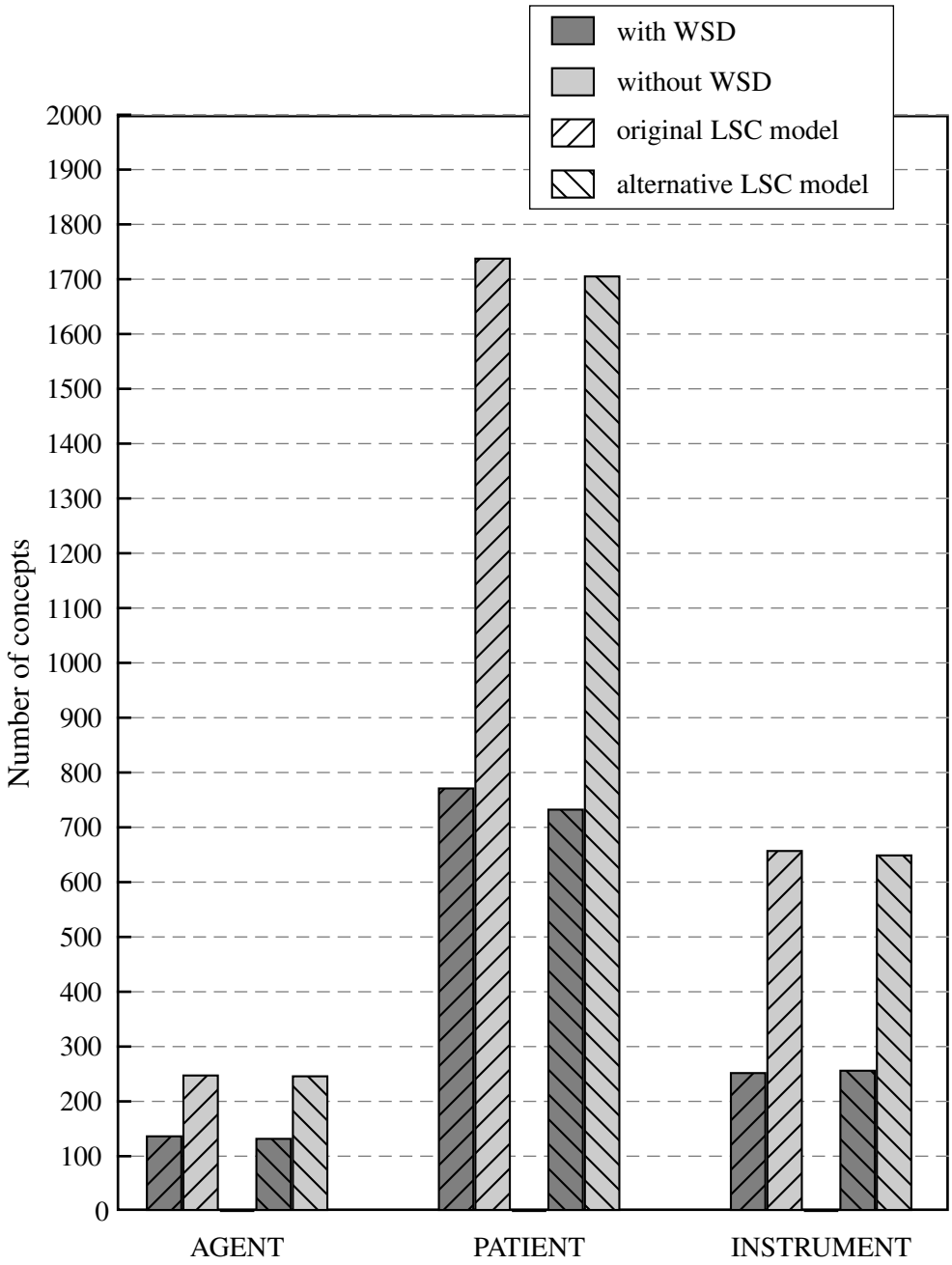


Figure 7.17: Average number of acquired concepts per verb for AGENT, PATIENT, and INSTRUMENT with the original and an alternative LSC model

role type	AGENT			
C	1 000 000			
gold standard noun concepts which are...	number (percentage)			
	with WSD		without WSD	
exactly matched	101	(13.8%)	232	(20.8%)
matched by 1 level hyponym	51	(6.9%)	103	(9.2%)
matched by 1 level hyperonym	68	(9.3%)	109	(9.8%)
matched by $\geq$ 2 level hyponym	13	(1.8%)	11	(1.0%)
matched by $\geq$ 2 level hyperonym	215	(29.3%)	311	(27.8%)
not matched	286	(39.0%)	351	(31.4%)
$\sum$	734		1117	
acquired noun concepts per verb	133.5		247.6	

Table 7.26: Results for the AGENT role with an alternative LSC model

role type	PATIENT			
C	1 000 000			
gold standard noun concepts which are...	number (percentage)			
	with WSD		without WSD	
exactly matched	89	(31.1%)	184	(39.8%)
matched by 1 level hyponym	51	(17.8%)	85	(18.4%)
matched by 1 level hyperonym	29	(10.1%)	42	(9.1%)
matched by $\geq$ 2 level hyponym	21	(7.3%)	25	(5.4%)
matched by $\geq$ 2 level hyperonym	35	(12.2%)	60	(13.0%)
not matched	61	(21.3%)	66	(14.3%)
$\sum$	286		462	
acquired noun concepts per verb	734.1		1706.9	

Table 7.27: Results for the PATIENT role with an alternative LSC model

role type	INSTRUMENT			
C	1 000 000			
gold standard noun concepts which are...	number (percentage)			
	with WSD		without WSD	
exactly matched	55	(12.6%)	128	(17.7%)
matched by 1 level hyponym	24	(5.5%)	57	(7.9%)
matched by 1 level hyperonym	29	(6.6%)	46	(6.4%)
matched by $\geq$ 2 level hyponym	11	(2.5%)	9	(1.2%)
matched by $\geq$ 2 level hyperonym	36	(8.2%)	97	(13.4%)
not matched	282	(64.5%)	387	(53.5%)
$\sum$	437		724	
acquired noun concepts per verb	257.8		650.6	

Table 7.28: Results for the INSTRUMENT role using an alternative LSC model

ber of different LSC models (i.e. models with different numbers of classes and iterations) would be necessary.

7.7 The Impact of Virtual Leaves

In section 7.3.1, I explained the role that virtual leaves on the acquired tree cuts play for matching gold standard relations. The current section investigates the performance achieved without taking virtual leaves into account. But before turning to the actual experiments, I would like to prepend a general discussion about the proper treatment of virtual leaves in the context of our learning approach.

Consider again the example in section 7.3.1. Figure 7.4 on page 219 displays the tree cut model acquired for the Agent of <travel>. The gold standard relations for this verb concept are listed under (7.5). I stated above that all these relations are matched by virtual leaves on the tree cut: <traveler>, <voyager>, and <wanderer> are exactly or approximately hit by <REST::traveler#traveller>; <salesman> is matched by the over-general concept <REST::worker>. However, in a strict technical sense, these assertions are not correct. In the hierarchy augmented by virtual leaves, <REST::traveler> is an immediate hyponym of <traveler>. Hence, the former concept matches the latter only approximately, not exactly. Furthermore, <voyager>, and <wanderer> are not matched by <REST::traveler> at all, since all three concepts are immediate hyponyms of <traveler>. Likewise, <salesman> is not matched by <REST::worker>. Both are subconcepts of <worker>, but neither of them is a subconcept of the other. Indeed, for counting matches in the evaluation experiments, I treat a virtual leaf as if it were the underlying WordNet concept, e.g. I identify <REST::traveler> as <traveler>. In this sense, <REST::traveler> does match <traveler>, <voyager>, and <wanderer>. One could argue that this proceeding is technically inexact. However, it is justifiable for linguistic reasons.

Virtual leaves had to be introduced due to formal requirements of the tree cut approach: all word senses have to be represented as leaves of the hierarchy tree, while inner nodes represent abstract semantic classes. For example, <REST::worker> represents a word sense of "worker", whereas <worker> represents a semantic class that comprises all kinds of worker. However, the semantic difference between a concept and the corresponding virtual leaf is not quite clear. Indeed, the word sense of "worker" represented by <REST::worker> refers to a concept that captures any kind of worker—which is just the concept represented by <worker>. In the wordnet framework, synsets simultaneously represent both word senses and semantic classes. This design principle appears natural and has proved its worth in the numerous wordnet applications in the past years. Therefore, it is as well appropriate to abolish the technically motivated separation of these two functions for the extraction of role relations from tree cuts and equate a virtual leaf with its underlying "original" synset to this end.

But what does a preferred virtual leaf in a tree cut model imply? It means that the words with the corresponding sense are more likely to appear as a complement of the examined verb concept than by chance. <REST::worker> is acquired as a preferred Agent of <travel> because "worker" cooccurs with "travel" to a significant extent. According to the tree cut in figure 7.4, nouns referring to more specific senses (such as "journeyman") occur with that verb as well. Thus, the noun senses cooccurring with a certain verb are located at differing generalisation levels themselves. A (more or less) abstract virtual leaf in a tree cut model is no generalisation *from* the data, but reflects generalisation which is immediately manifest *in* the data. This basic phenomenon has been neglected in the discourse

role type	AGENT			
C	1 000 000			
gold standard noun concepts which are...	number (percentage)			
	with WSD		without WSD	
exactly matched	49	(6.6%)	114	(10.3%)
matched by 1 level hyponym	110	(14.9%)	215	(19.4%)
matched by 1 level hyperonym	22	(3.0%)	44	(4.0%)
matched by ≥ 2 level hyponym	19	(2.6%)	15	(1.4%)
matched by ≥ 2 level hyperonym	62	(8.4%)	115	(10.4%)
not matched	476	(64.5%)	605	(54.6%)
$\sum$	738		1108	
acquired noun concepts per verb	138.1		249.1	

Table 7.29: Results for the AGENT role without allowing matches of virtual leaves

about the tree cut approach so far. This discourse has focused on the MDL approach as a source of generalisation. Immediate evidence from the data is another source. In the experiments described below, I examine the the importance of this source.

These experiments differ from the basic performance tests in section 7.3 by ignoring virtual leaves when comparing the learned cuts with the gold standard. The results for AGENT, PATIENT, and INSTRUMENT are shown in tables 7.29 to 7.31. The differences to the basic performance results exhibit a similar pattern for all three role types. I will exemplify this pattern by comparing the results for AGENT without WSD (cf. figure 7.18). First of all, disregarding virtual leaves leads to a decrease of approximate accuracy: 33.7% vs. 39.9% (in figure 7.18, the respective sum of the first three columns). It is striking that while the rate of exact matches drops significantly (10.3% vs. 20.9%), the rate of matches by 1-level hyponyms increases (19.4% vs. 9.2%). Obviously, in many cases concepts which are one level too specific are present in addition to an exactly matching virtual leaf; those concepts compensate the loss of that leaf (e.g. in the cut in figure 7.4, the loss of <REST::traveler> would be compensated by the presence of concepts like <passenger> or <guest>, which match <traveler> at least approximately). The matching rate of hyperonym concepts drops dramatically, especially for two or more levels up in the hierarchy (10.4% vs. 28.2%). This mainly accounts for a comparably dramatic increase of relations not matched at all (54.6% vs. 31%). Overall, one can conclude that virtual leaves play a significant role for acquiring thematic relations.

7.8 The Impact of the Generalisation Level

In section 7.3, I pointed out that a considerable amount of relations in the gold standard involve very specific noun concepts. I presented examples which demonstrate that these specific concepts introduce a sparse data problem: often they are missed by the acquired tree cuts because the respective words do not occur in the data. I claimed that this problem is primarily responsible for the rather moderate performance for AGENT and INSTRUMENT in comparison to PATIENT. To verify this claim and to test the significance of this data sparseness problem, I did evaluation experiments in which the gold

role type	PATIENT			
C	1 000 000			
gold standard noun concepts which are...	number (percentage)			
	with WSD		without WSD	
exactly matched	34	(11.9%)	63	(13.6%)
matched by 1 level hyponym	99	(34.7%)	205	(44.4%)
matched by 1 level hyperonym	9	(3.2%)	14	(3.0%)
matched by $\geq$ 2 level hyponym	36	(12.6%)	34	(7.4%)
matched by $\geq$ 2 level hyperonym	1	(0.4%)	8	(1.7%)
not matched	106	(37.2%)	138	(29.9%)
$\sum$	285		462	
acquired noun concepts per verb	772.8		1739.4	

Table 7.30: Results for the PATIENT role without allowing matches of virtual leaves

role type	INSTRUMENT			
C	1 000 000			
gold standard noun concepts which are...	number (percentage)			
	with WSD		without WSD	
exactly matched	23	(5.8%)	61	(8.6%)
matched by 1 level hyponym	38	(9.6%)	110	(15.5%)
matched by 1 level hyperonym	7	(1.8%)	15	(2.1%)
matched by $\geq$ 2 level hyponym	13	(3.3%)	15	(2.1%)
matched by $\geq$ 2 level hyperonym	5	(1.3%)	2	(0.3%)
not matched	308	(78.2%)	506	(71.4%)
$\sum$	394		709	
acquired noun concepts per verb	253.5		658.7	

Table 7.31: Results for the INSTRUMENT role without allowing matches of virtual leaves

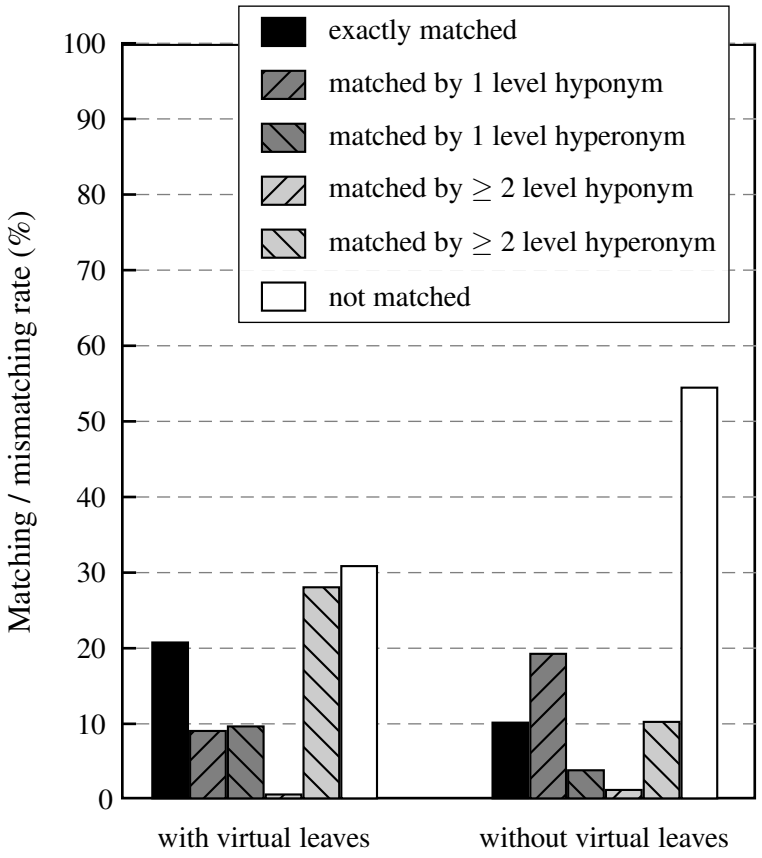


Figure 7.18: Matching / mismatching rates for AGENT (without WSD) with and without taking into account virtual leaves

standard was restricted to relations at a comparably high level of generalisation. More precisely, I fixed a certain abstraction level in the WordNet hierarchy and only employed gold standard relations where the noun concepts do not fall below this level.

First of all, I had to define this limiting abstraction level. It is inevitable that such a definition has to be arbitrary to some extent. However, to get an idea about the performance w.r.t. general concepts, the *exact* course of the border which separates concepts with sufficient generality from other concepts is secondary. A straightforward way of fixing the limit would be to start from the WordNet root nodes and stipulate that all concepts located at 0–n levels below these nodes are considered as sufficiently general. However, this would neglect the fact that for the individual role types there are certain effective upper bounds of abstraction, which have to be taken into account. These upper bounds correspond to the semantic filters specified in section 6.5. For example, the INSTRUMENT filter includes (among others) the concepts <inanimate_object> and <body_part>. This does not only constrain the semantic range of the INSTRUMENT role, but also implies that INSTRUMENT concepts should not be located above these two nodes. For AGENT, the same holds with the concepts <life_form> and <causal_agent>.[4] For PATIENT, there is no semantic restriction. I decided to define a common limit of generality for all role types which takes into account these upper bounds. As a starting point, I adopted all WordNet roots plus all "non-root" concepts in the semantic filters. This yields the following list:

<life_form#organism#being#living_thing>
<causal_agent#cause#causal_agency>
<object#inanimate_object#physical_object>
<body_part>
<part#piece>[5]
<entity>
<psychological_feature>
<abstraction>
<location>
<shape#form>
<state>
<event>
<act#human_action#human_activity>
<group#grouping>
<possession>
<phenomenon>

I consider a concept as sufficiently general if it is located at most 2 levels below one of the concepts in the list. The obtained abstraction limit covers the top levels of the complete hierarchy (via the root nodes) on the one hand and allows further specificity according to semantic filter definitions on the other hand.

The experiments restricted to general gold standard relations defined in this way yield interesting results. Table 7.32 shows the matching rates for AGENT. Figure 7.19 provides a bar diagram of the accuracy rates and the complete mismatches. The rate of exact or approximate matches is almost doubled (without WSD, 72.8% as opposed to 39.9% basic performance). Conversely, the rate of com-

[4]Note that the other concepts included in the filters of AGENT and INSTRUMENT are WordNet roots.

[5]This concept is the hyperonym of <body_part> and an immediate hyponym of <entity>. I included it to obtain a smoother range of general concepts.

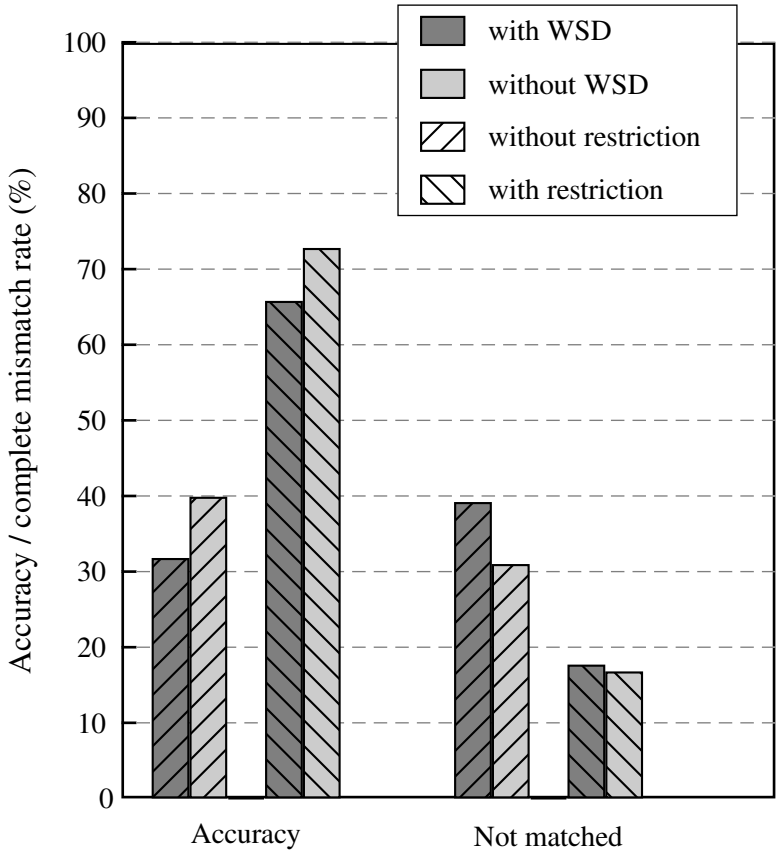


Figure 7.19: Accuracy (approximate match rates) and complete mismatch rates for AGENT with and without restriction to relations at higher generalisation levels

role type	AGENT	
C	1 000 000	
gold standard noun concepts which are...	number (percentage) with WSD	without WSD
exactly matched	14 (17.7%)	34 (27.2%)
matched by 1 level hyponym	27 (34.2%)	38 (30.4%)
matched by 1 level hyperonym	11 (13.9%)	19 (15.2%)
matched by $\geq$ 2 level hyponym	3 (3.8%)	2 (1.6%)
matched by $\geq$ 2 level hyperonym	10 (12.7%)	11 (8.8%)
not matched	14 (17.7%)	21 (16.8%)
$\sum$	79	125

Table 7.32: Results for the AGENT role restricted to relations at higher generalisation levels

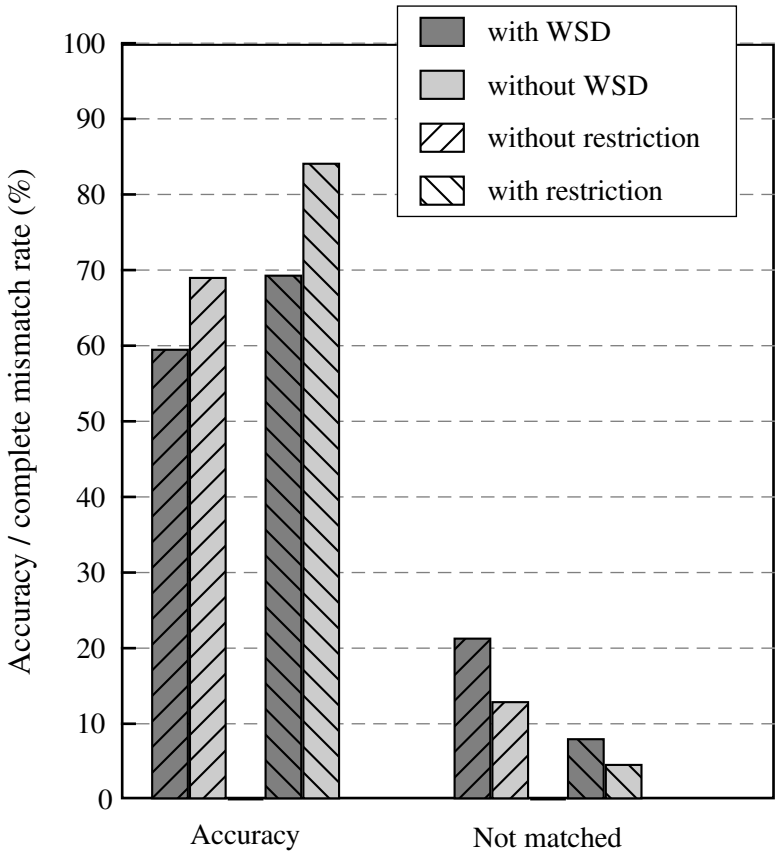


Figure 7.20: Accuracy (approximate match rates) and complete mismatch rates for PATIENT with and without restriction to relations at higher generalisation levels

pletely unmatched relations falls below the respective rate in the basic experiment to a considerable extent (without WSD, 16.8% as opposed to 31%). Thus, the learning algorithm performs considerably better when restricted to general relations.

Table 7.33 displays the accuracy rates for PATIENT. Figure 7.20 provides a corresponding bar diagram. As for AGENT, the rates of exact or approximate matches significantly exceed the rates in the basic experiments (without WSD, 84.2% vs. 69.1%). Furthermore, the rate of unmatched concepts is much lower (without WSD, 4.7% vs. 13%). However, the difference in accuracy is not so dramatic as for AGENT. In fact, the results for PATIENT are already quite satisfying in the basic setting where all gold standard relations are taken into account. In section 7.3.2, I mentioned findings which indicate that the relative portion of general relations in the gold standard is higher for PATIENT than for AGENT. The results displayed here confirm this claim. Without WSD, 125 general AGENT relations are taken into account. In the basic experiment, 1108 relations are considered. This means that only 11% of the relations employed to measure basic performance fall under our definition of general relations. For PATIENT, the corresponding numbers of relations are 107 (general) and 462 (all). Thus, around a quarter (23%) of PATIENT relations used to test basic performance are general. In other

role type	PATIENT			
C	1 000 000			
gold standard noun concepts which are...	number (percentage)			
	with WSD		without WSD	
exactly matched	21	(33.9%)	45	(42.1%)
matched by 1 level hyponym	18	(29.0%)	40	(37.4%)
matched by 1 level hyperonym	4	(6.5%)	5	(4.7%)
matched by $\geq$ 2 level hyponym	14	(22.6%)	10	(9.3%)
matched by $\geq$ 2 level hyperonym	0	(0.0%)	2	(1.9%)
not matched	5	(8.1%)	5	(4.7%)
$\sum$	62		107	

Table 7.33: Results for the PATIENT role restricted to relations at higher generalisation levels

role type	INSTRUMENT			
C	1 000 000			
gold standard noun concepts which are...	number (percentage)			
	with WSD		without WSD	
exactly matched	3	(17.6%)	10	(38.5%)
matched by 1 level hyponym	4	(23.5%)	6	(23.1%)
matched by 1 level hyperonym	0	(0.0%)	2	(7.7%)
matched by $\geq$ 2 level hyponym	1	(5.9%)	0	(0.0%)
matched by $\geq$ 2 level hyperonym	0	(0.0%)	1	(3.8%)
not matched	9	(52.9%)	7	(26.9%)
$\sum$	17		26	

Table 7.34: Results for the INSTRUMENT role restricted to relations at higher generalisation levels

words, the portion of general relations, which are matched by the algorithm with more reliability, is considerably higher. This provides an explanation for the fact that the basic results for PATIENT are better than for AGENT.

Table 7.34 shows the matching rates for INSTRUMENT. These rates indicate that also for this role type, the results significantly exceed the basic performance. However, the number of relations taken into account (17 or 26, respectively) show that the portion of general INSTRUMENT relations in the gold standard is too small to draw reliable conclusions. This is a remarkable finding: Obviously, the overwhelming majority of INSTRUMENT relations contain rather specific concepts. This explains why the performance for INSTRUMENT is even poorer than for AGENT.

To exclude the possibility that the arbitrary choice of the limit of abstraction inadequately biases the results, I repeated the experiments extending the range of general concepts, i.e. I permitted concepts located up to 3 or 4 (rather than 2) levels below the concepts listed above. These experiments confirm the described tendencies. With each further included level, the accuracy rates decrease and, naturally, the numbers of relations taken into account increase. In any case, the portion of captured relations is

significantly higher for PATIENT than for AGENT, and remains fairly low for INSTRUMENT.

7.9 Summary and Conclusion

The detailed evaluation presented in this chapter revealed some interesting insights into the performance of the role acquisition approach on the one hand and the nature and the suitability of the gold standard on the other hand. In detail, the following findings concerning the learning approach can be observed:

- As the experiments show, the approach exhibits a good performance of learning role relations involving noun concepts at higher levels of abstraction.
- On the other hand, the performance of matching gold standard relations at low generalisation levels was rather poor. This is essentially due to a sparse data problem: often the specific noun concepts in the gold standard do not occur (or, more precisely, do not co-occur with the respective verb in question) in the data. Therefore, it would be unreasonable to conclude that the approach is not capable of appropriately acquiring role relations with specific noun concepts. Rather, all that follows from the experiments is that this capability cannot be assessed with a gold standard as employed here. For example, as we have seen in section 7.3.1, the gold standard contains the relation <rescue> INVOLVED_AGENT <liberator>. The tree cut acquired for the Agent of <rescue> (cf. figure 7.3 on page 218) misses the concept <liberator>, but comprises a number of concepts which as well refer to typical instances of rescuers, e.g. <fireman> or <policeman>. It is mere coincidence that the gold standard includes <liberator> but not <fireman>, or that nouns like "fireman" occur in the data but "liberator" does not.
- Applying semantic filters reduces the number of acquired relations to a large extent, accompanied by a tolerable decrease of accuracy. This is very useful in the semi-automatic scenario where the learned relations undergo manual post-editing.
- Employing argument clustering for linking instead of simple heuristics neither enhances nor lowers accuracy. However, an overall improvement of the coverage of the gold standard could be observed, i.e. the range of captured relations increases with argument clustering. This improvement, though, is rather moderate.
- My experiments yield no evidence that the variation of parameters of the LSC model which underlies the learning approach does not have a crucial influence on performance.
- The inclusion of virtual leaves improves accuracy, which indicates that they constitute a valuable source of information for learning role relations. Preferred virtual leaves in a tree cut model capture (different levels of) generalisation immediately present *in* the corpus data. This complements the ability of the MDL-based algorithm to generalise *from* the data.

The following findings and conclusions concern the proper application of a gold standard in general as well as specific properties of the actual gold standard originating from the EWN database in particular:

- For the compilation of a gold standard for evaluating approaches for learning role relations, only relations involving noun concepts at higher abstraction levels should be taken into account. Relations with a noun concept at low generalisation levels impose a severe sparse data problem.

- According to this condition, the suitability of the set of role relations in EWN differs for the different role types.
  - For AGENT, there is a sufficient number of relations with higher generalisation. Such relations can be extracted and employed as a gold standard where the exact limit of "sufficient abstraction" adopted is secondary.
  - For PATIENT, the portion of relations with suitable generality seems to be so dominant that satisfying accuracy rates are achieved even without any preselection of general relations. However, such a preselection improves the results for this role type as well.
  - Although the number of INSTRUMENT relations in EWN is quite high, there are hardly any relations at suitable abstraction levels. Thus, the information encoded in EWN does not allow the compilation of a gold standard for INSTRUMENT.
- For LOCATION and directional role types, the amount of relations in EWN is not sufficient to serve as a gold standard. Although the absolute number of LOCATION relations appears adequate, the majority of involved verb concepts do not or do only sparsely occur in the data.

From these points it follows that the EWN database allows the extraction of a gold standard only for PATIENT and (to a limited extent) for AGENT.

Chapter 8

Conclusion

This chapter provides a concise summary of this thesis by highlighting its main results. Furthermore, it mentions several directions for future research. The task of this thesis has been to devise a strategy of learning thematic role relations for lexical-semantic nets, in particular wordnets. To fulfil this task, I have developed an approach that takes as input a syntactically analysed corpus and WordNet, and employs probabilistic techniques and linguistically motivated heuristics to learn role relations. These candidate relations are subject to manual inspection before being integrated into the wordnet to be extended.

The approach comprises the following components:

- The central module is a **learning algorithm that acquires selectional preferences** from corpus data. This approach is an enhancement of a method proposed in (Abe & Li 1996). This method acquires a tree cut model to capture the selectional preferences of a verb for a certain argument. A tree cut model is a horizontal cut through the WordNet noun hierarchy where the concepts located on the cut are assigned numerical preference values; a value > 1 indicates preference, a value < 1 indicates dispreference. The location of the tree cut in the hierarchy, and thus the generalisation level of the acquired selectional preferences, is determined by the theoretically well-founded MDL principle. The general approach of (Abe & Li 1996) is described in section 3.4.3. However, preliminary experiments showed that this approach exhibits a highly undesired behaviour regarding the task of this thesis. The acquired generalisation level inadequately depends on the frequency of the respective verb: the lower the frequency, the higher the generalisation level. Section 4.1 describes these preliminary experiments and discusses this behaviour. I proposed a modification of the method that overcomes this drawback. This modification deviates from the MDL principle in its strict sense, but still conforms to the Bayesian Learning paradigm. The modified approach I propose is described in section 4.2. This section also reports first tests which indicate its appropriateness.
- The tree cut approach learns selectional preferences by generalising over nouns occurring as a certain argument of a certain verb. Originally, this approach—as well as comparable approaches introduced in chapter 3—learns selectional preferences for *syntactic arguments*, since the available training data (parsed corpora) only indicate syntactic complements. However, the task of this thesis requires to learn preferences for thematic roles, i.e. *semantic arguments*. For this reason, I developed a **linking strategy** to map syntactic complements to their underlying roles. This allows acquiring selectional preferences by generalising over nouns that occur as the filler

of a certain thematic role of the examined verb. This linking strategy (described in chapter 6) consists of three stages. The first stage computes clusters of semantically similar syntactic relations between verbs and nouns. The second stage applies linguistically motivated heuristics for assigning appropriate thematic role types to these clusters. These stages yield thematic role relations between verbs and nouns, which are—after undergoing certain preprocessing steps (see below)—fed into the preference acquisition algorithm. A third stage of linking consists of semantic filters which are applied to the acquired tree cuts to narrow down the space of candidate role relations.

- The training data contain non-disambiguated verb and noun forms that have to be mapped to the corresponding WordNet concepts before the learning algorithm can be applied. This is done in two preprocessing steps. Firstly, the word forms in the data are (partially) lexically disambiguated. For this step, I employed a **WSD strategy** that combines semantic clustering with a method which disambiguates words within a cluster (cf. section 5.2). Secondly, the frequencies of the noun senses retrieved by this step are propagated to their hyperonym concepts in the WordNet hierarchy. This propagation has to take into account the fact that WordNet is not a pure tree, but a DAG, i.e. a concept may have multiple immediate hyperonyms. In section 5.3, I propose a principled solution to that so-called **DAG issue**.

Figures 8.1 and 8.2 show a data flowchart which provides an overview of the overall approach developed in this thesis. This flowchart (to be read top down) makes explicit the dependencies between the different processing modules (represented by rectangles) and the different types of input and intermediary data (represented by parallelograms, in some cases associated with a comment comprising a formula which explicates the respective kind of data). While figure 8.1 depicts the preprocessing phase including the role assigning strategy and the mapping of word forms to WordNet concepts, figure 8.2 displays the preference acquisition step itself (taking into account a particular role) and the postprocessing step of semantic filtering. Both sub-flowcharts are associated via a connector symbol labelled with "A".

In section 1.3, I stated two desiderata for the acquisition approach: it should be language-independent and open to different linguistic theories of thematic roles. These desiderata are met to the largest extent possible. Almost all components mentioned above employ statistical methods which are independent from particular languages and thematic role theories. There are only two exceptions, namely stage 2 and 3 of the linking strategy. Stage 2 (heuristics for assigning role types to verb–noun clusters) makes heavy use of linguistic findings about diathesis alternations in English and the inventory and definition of role relations in EWN. Thus, this module does not satisfy either of the two desiderata. Stage 3 (semantic filters) makes use of general semantic characterisations of role types (e.g. that Agents are usually animate), thus violating the desideratum of theory-neutrality. Applying my approach for another language or under other theoretical assumptions requires the reformulation of these two modules. The question whether it is possible to design them in a universal way is intrinsically connected to the question whether and to what degree linking follows universal principles, and how these principles look like. This question is subject to linguistic research and far from being conclusively answered.

The approach presented here is not limited to the task of acquiring role relations for wordnets. In principle, it is suitable to enrich other lexical resources with selectional restrictions as well. In particular, it offers the flexibility to globally influence the level of generalisation of the acquired preferences by the choice of the constant C. In this way, it can be fine-tuned to the needs of different applications. For example, selectional restrictions for a lexicon supplementing a parser may be at a rather abstract level, while preferences encoded in a knowledge base for semantic and pragmatic inferencing should

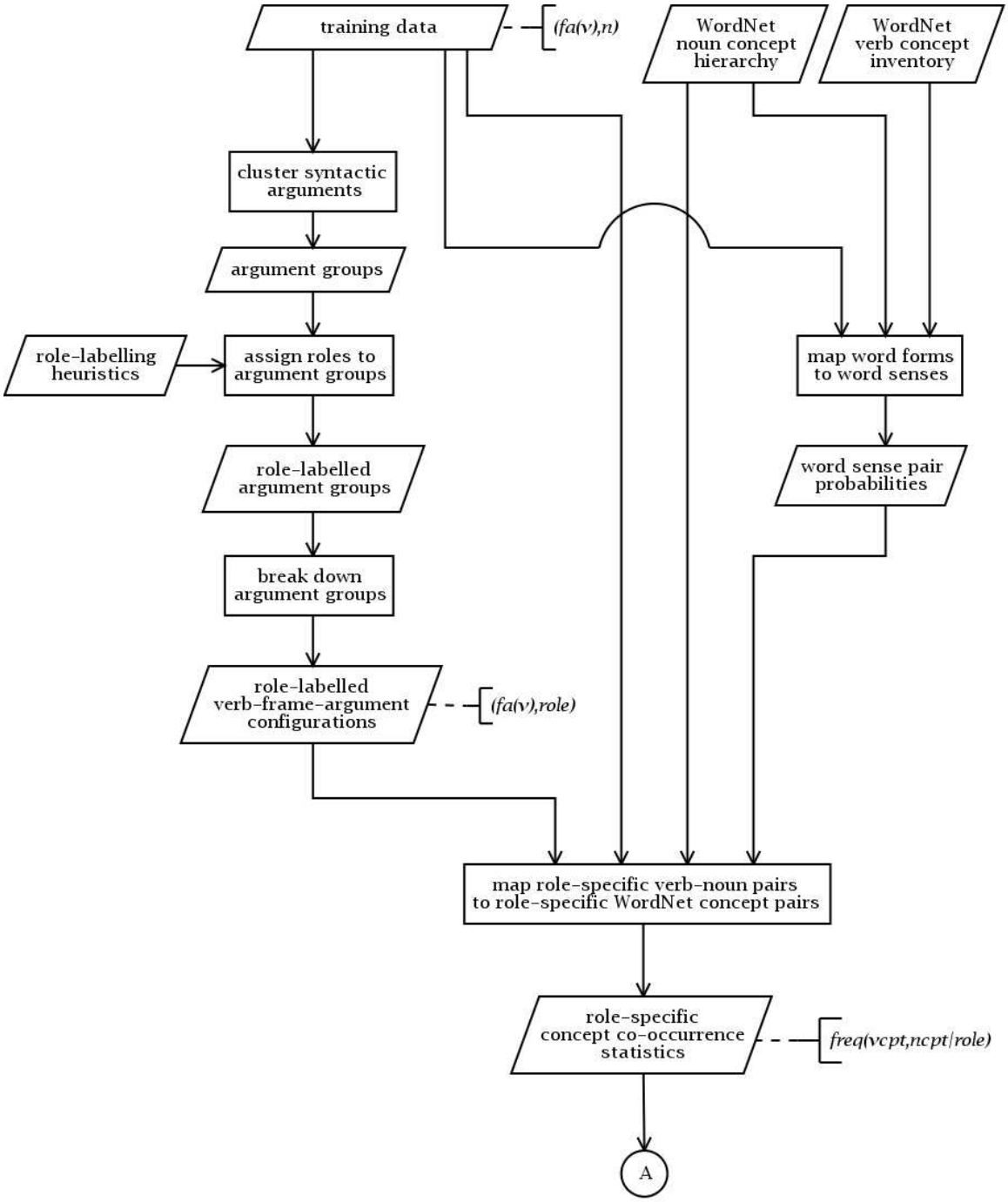


Figure 8.1: Part 1 of the data flowchart giving an overview of the approach developed in this thesis

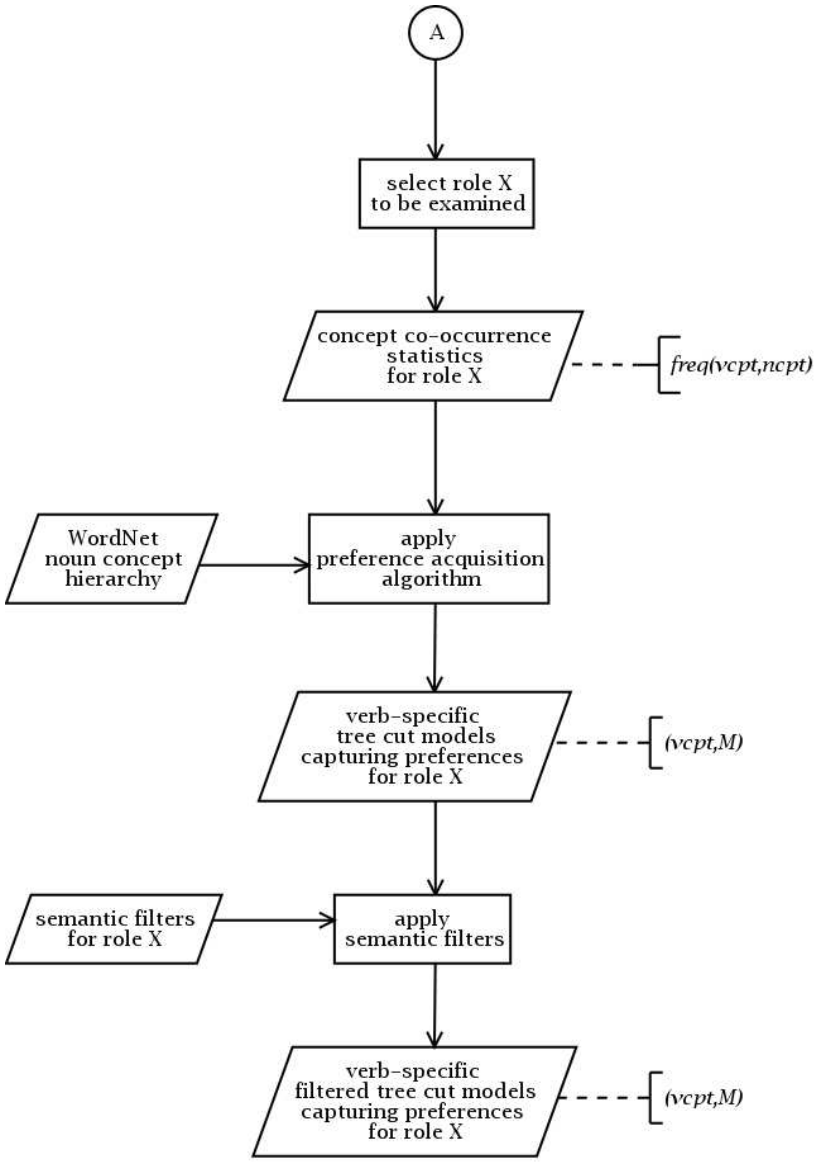


Figure 8.2: Part 2 of the data flowchart giving an overview of the approach developed in this thesis

be rather specific. My approach offers the possibility of respective adjustments.

A further important issue in this work has been to devise a principled evaluation procedure for the approach developed here and related methods. To this end, I extracted a **gold standard** from the data in EuroWordNet (cf. section 5.4) and proposed a **setup and criteria for evaluation** (cf. section 7.2). Chapter 7 reports a detailed evaluation of different aspects of my role acquisition approach. Overall, the results (accuracy rates of up to 84%) are satisfying and show that the approach works effectively. This evaluation also assesses the general feasibility of employing a gold standard for evaluating the task of this thesis, and reveals the strengths and weaknesses of the particular gold standard retrieved from EWN. In general, a gold standard is only suitable to evaluate the acquisition of selectional preferences located at higher generalisation levels. For lower abstraction levels, a sparse-data problem makes a fair evaluation impossible. The particular gold standard used in this work is suitable to evaluate the acquisition of AGENT and PATIENT relations.

Apart from the main focus of this work, the thesis also contains some general considerations regarding diverse theoretical issues:

- In section 5.2.1, I present arguments questioning a widely-held assumption that is often referred to when estimating word sense frequencies from corpora which are not lexically disambiguated. This assumption, which I call the uniformity hypothesis, states that, if further information concerning sense distributions of word forms are missing, it is suitable to stipulate for each word form a uniform probability distribution of its senses. Thus, if a word form with n senses has the frequency f, each of its senses is assigned the frequency $\frac{f}{n}$. I have pointed out a general shortcoming of this approach, related to the phenomenon of regular polysemy, and argued for the importance of employing (at least partial) word sense disambiguation.

- In section 4.1.2, I discuss the class of tree cut models from the perspective of the MDL principle. The undesired behaviour yielded if MDL is applied to select a tree cut model—the lower the frequency of a verb, the more general the selected tree cut—is a consequence of an inherent property of MDL. This property can be stated as follows: the more data have to be described, the higher the bias towards a more complex model, i.e. a model with many parameters. I argue that this behaviour is perfectly adequate for many model classes addressed in the MDL literature (e.g. decision trees), but inappropriate for the class of tree cut models. In this context, I introduce the distinction between the complexity of a model and the extent to which this model explicitly expresses regularities which are implicit in the data. I hope that this discussion may provide a deeper insight into the MDL principle and the conditions of its application.

- In section 6.2.2, I provide a linguistic interpretation of the latent semantic clustering (LSC) approach developed by (Rooth et al. 1998), which I employ in my linking strategy. I argue that an LSC model can be viewed as a probabilistic implementation of Dowty's theory of thematic roles. Note that this statement does not conflict with the desideratum of theory-neutrality. Interpreting the LSC approach in terms of Dowty's approach is evident, but not compulsory.

These considerations should be understood as contributions to the discussion of the respective issues.

At the end of this thesis, I would like to briefly indicate some possible directions of future research. First of all, it would be interesting to assess whether changes in individual modules would improve the performance of the overall approach. Such changes could be, among others:

- *Preference/dispreference borderline:* The distinction between preferred and dispreferred concepts in the tree cut model is made according to the respective preference values. Currently, the theoretically justified value 1 serves as the border between preference and dispreference. Only those concepts that are above that borderline are regarded as preferred, the others are regarded as dispreferred and discarded. One could arbitrarily choose a different value for that border. A higher limit would further narrow down the set of acquired candidate relations, resulting in less effort for manual inspection. However, a certain portion of appropriate relations would be missed. A lower limit could possibly capture more relations that are appropriate, but the candidate set would grow. These effects should be investigated empirically.

- *Word sense disambiguation:* Instead of the WSD approach I use, any existing WSD approach related or adaptable to WordNet could be employed. The better the disambiguation module, the better the overall performance.

Another research direction concerns the linking strategy. This approach could be of interest independently from acquiring thematic role relations. It could be adapted and employed for related tasks such as the detection of diathesis alternations and automatic verb classification in general (cf. section 6.6). Therefore, this approach deserves further investigation. A first step could be a systematic evaluation. In this work, the linking strategy was evaluated indirectly by employing it for a concrete task, namely role relation acquisition. One could imagine a complementary setup in which the acquisition of roles is employed to evaluate linking. Such a setup could be designed as follows: For a verb occurring in the gold standard, the learning algorithm acquires tree cut models from the respective noun sets which are determined by argument grouping via LSC clusters. The resulting tree cuts are compared to the gold standard relations of a certain role type (e.g. AGENT) of the examined verb; the cut which is most similar to these relations is assigned the corresponding role type. Then one evaluates whether or not the linking heuristics would assign this same role type to the clusters which underly that tree cut.

Finally, the overall acquisition approach has been evaluated by comparison to a gold standard. It would be interesting to perform additional evaluations by employing the approach for concrete NLP applications, such as parsing, WSD, or information extraction. In particular, one could test the impact of the choice of C, i.e. the global abstraction bias, on the performance of the respective applications.

In any case, much work remains to be done.

Bibliography

Abe, Naoki & Hang Li (1996), Learning word association norms using tree cut pair models, *in* 'Proc. of 13th Int. Conf. on Machine Learning', Bari.

Abney, Steven (1996*a*), Partial parsing via finite-state cascades, *in* J.Carroll, ed., 'Workshop on Robust Parsing (ESSLLI '96)', Prague, pp. 8–15.

Abney, Steven (1996*b*), Statistical Methods and Linguistics, *in* J.Klavans & P.Resnik, eds, 'The Balancing Act', MIT Press, Cambridge, MA.

Abney, Steven & Marc Light (1998), Hiding a semantic class hierarchy in a Markov model. Unpublisched manuscript.

Abney, Steven & Marc Light (1999), Hiding a semantic hierarchy in a Markov model, *in* 'Proc. of ACL'99 Workshop on Unsupervised Learning in Natural Language Processing', College Park, MD, pp. 1–8.

Agirre, Eneko & David Martinez (2002), Integrating selectional preferences in WordNet, *in* 'Proc. of First International WordNet Conference', Mysore.

Alonge, Antonietta, ed. (1996), *Definition of the links and subsets for verbs*, EuroWordNet (LE2-4003). Deliverable D006.

Alshawi, Hiyan, ed. (1992), *The Core Language Engine*, MIT Press, Cambridge, MA.

Biemann, Chris, Stefan Bordag & Uwe Quasthoff (2004), Automatic acquisition of paradigmatic relations using iterated co-occurrences, *in* 'Proc. of LREC 2004', Lisboa.

Bierwisch, Manfred (1983), Semantische und konzeptuelle Repräsentation lexikalischer Einheiten, *in* R.Ružička & W.Motsch, eds, 'Untersuchungen zur Semantik', Vol. XXII of *studia grammatica*, Akademie-Verlag, Berlin, pp. 61–99.

Boguraev, Branimir & James Pustejovsky, eds (1996), *Corpus Processing for Lexical Acquisition*, MIT Press, Cambridge, MA.

Budanitsky, Alexander & Graeme Hirst (2001), Semantic distance in WordNet: An experimental, application-oriented evaluation of five measures, *in* 'Proc. of WordNet and Other Lexical Resources Workshop, NAACL', Pittsburgh.

Buenaga Rodriguez, Manuel de, Jose Maria Gomez-Hidalgo & Belen Diaz-Agudo (1997), Using WordNet to complement training information in text categorization, *in* 'Proc. of RANLP-97', Tzigov Chark.

Buitelaar, Paul (1998), CoreLex: An ontology of systematic polysemous classes, *in* 'Proc. of FOIS98, International Conference on Formal Ontology in Information Systems', Trento.

Buitelaar, Paul (2000), Reducing lexical semantic complexity with systematic polysemous classes and underspecification, *in* 'Proc. of ANLP2000 Workshop on Syntactic and Semantic Complexity in Natural Language Processing Systems', Seattle.

Burnard, Lou (1995), *Users Reference Guide for the British National Corpus. Version 1.0*, Oxford University Computing Services.

Burns, Kathy J. & Anthony R. Davis (1999), Building and maintaining a semantically adequate lexicon using CYC, *in* E.Viegas, ed., 'Breadth and Depth of Semantic Lexicons', Kluwer Academic Publishers, pp. 121–143.

Carreras, Xavier & Lluís Màrques (2004), Introduction to the CoNLL-2004 shared task: Semantic role labeling, *in* 'Proc. of CoNLL-2004', Boston, MA, pp. 89–97.

Carroll, Glenn & Mats Rooth (1998), Valence induction with a head-lexicalized PCFG, *in* 'Proc. of EMNLP-3', Granada.

Charniak, Eugene (1993), *Statistical Language Learning*, MIT Press, Cambridge, MA.

Chomsky, Noam (1965), *Aspects of the Theory of Syntax*, MIT Press, Cambridge, MA.

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

Clark, Stephen & David Weir (2002), 'Class-based probability estimation using a semantic hierarchy', *Computational Linguistics* **28**(2), 187–206.

Cover, Thomas M. & Joy A. Thomas (1991), *Elements of Information Theory*, John Wiley & Sons, New York.

Dagan, Ido, Lillian Lee & Fernando Pereira (1997), Similarity-based methods for word sense disambiguation, *in* 'Proc. of ACL 35 / EACL 8', San Francisco, pp. 56–63.

Dagan, Ido, Lillian Lee & Fernando Pereira (1999), 'Similarity-based models of cooccurrence probabilities', *Machine Learning* **34**(1–3), 43–69.

de Marcken, Carl G. (1996), Unsupervised Language Acquisition, PhD thesis, MIT.

Dempster, Arthur, Nan Laird & Donald Rubin (1977), 'Maximum likelihood from incomplete data via the EM algorithm', *Journal of the Royal Statistical Society* **39**(B), 1–38.

Dowty, David (1991), 'Thematic proto-roles and argument selection', *Language* **67**(3), 547–619.

Drange, Theodore (1966), *Type Crossings. Sentential Meaninglessness in the Border Area of Linguistics and Philosophy*, Mouton & Co., The Hague.

Fass, Dan, James Martin & Elizabeth Hinkelman, eds (1992), *Computational Intelligence. Special Issue on Non-Literal Language*, Vol. 8.

Fellbaum, Christiane (1990), English verbs as a semantic net, *in* G.Miller et al., eds, 'Five papers on WordNet', number 43 *in* 'CSL Report', Cognitive Science Laboratory, Princeton University, pp. 40–61.

Fellbaum, Christiane, ed. (1998), *WordNet: An Electronic Lexical Database*, MIT Press, Cambridge, MA.

Fillmore, Charles J. (1968), The case for case, *in* E.Bach & R.Harms, eds, 'Universals in Linguistic Theory', Holt, Rinehart & Winston, New York, pp. 1–90.

Fillmore, Charles J. (1977), The case for case reopened, *in* P.Cole & J.Sadock, eds, 'Syntax and Semantics 8: Grammatical Relations', Academic Press, New York, pp. 59–82.

Fillmore, Charles J., Charles Wooters & Collin F. Baker (2001), Building a large lexical databank which provides deep semantics, *in* 'Proc. of Pacific Asian Conference on Language, Information and Computation', Hong Kong.

Fillmore, Charles J. & Collin F. Baker (2001), Frame semantics for text understanding, *in* 'Proc. of WordNet and Other Lexical Resources Workshop, NAACL', Pittsburgh.

Gildea, Daniel & Daniel Jurafsky (2002), 'Automatic labeling of semantic roles', *Computational Linguistics* **28**(3), 245–288.

Gonzalo, Julio, Felisa Verdejo, Irina Chugur & Juan Cigarran (1998), Indexing with WordNet synsets can improve text retrieval, *in* 'Proc. of COLING/ACL'98 Workshop on Usage of WordNet for NLP', Montréal.

Grishman, Ralph & Beth Sundheim (1996), Message Understanding Conference - 6: A brief history, *in* 'Proc. of COLING', Copenhagen.

Grishman, Ralph, Catherine Macleod & John Sterling (1992), Evaluating parsing strategies using standardized parse files, *in* 'Proc. of 3rd ACL Conference on Applied Natural Language Processing', Trento, pp. 156–161.

Gruber, Jeffrey (1965), Studies in Lexical Relations, PhD thesis, Indiana University Linguistics Club.

Hamp, Birgit & Helmut Feldweg (1997), GermaNet - a lexical-semantic net for German, *in* 'Proc. of ACL/EACL-97 Workshop on Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications', Madrid, pp. 9–15.

Hobbs, Jerry R. (1978), 'Resolving pronoun references', *Lingua* **44**, 311–338.

Huyck, Christian R. (2000), A practical system for human-like parsing, *in* 'Proc. of ECAI 2000', Berlin, pp. 436–440.

Ikegami, Yoshihiko (1993), Semantic analysis and the Activator, *in* 'Longman Language Activator', Longman Group, Harlow, UK, pp. F20–F21.

Jackendoff, Ray (1987), 'The status of thematic relations in linguistic theory', *Linguistic Inquiry* **18**(3), 369–411.

Jackendoff, Ray (1990), *Semantic Structures*, MIT Press, Cambridge, MA.

Jiang, Jay J. & David W. Conrath (1997), Semantic similarity based on corpus statistics and lexical taxonomy, *in* 'Proc. of International Conference on Research in Computational Linguistics ROCLING X)', Taipei.

Johnson-Laird, Philip N. (1983), How is the meaning of a word mentally represented?, *in* 'Mental Models: Towards a Cognitive Science of Language, Inference, and Consciousness', Harvard University Press, Cambridge, MA, pp. 205–242.

Katz, Jerrold J. & Jerry A. Fodor (1964), The structure of a semantic theory, *in* J.Fodor & J.Katz, eds, 'The Structure of Language. Readings in the Philosophy of Language', Prentice-Hall, Englewood Cliffs, NJ, pp. 479–518.

Killgarriff, Adam (1998), SENSEVAL: An exercise in evaluating word sense disambiguation programs, *in* 'Proc. of LREC 1998', Granada, pp. 581–588.

Kingsbury, Paul, Martha Palmer & Mitch Marcus (2002), Adding semantic annotation to the Penn TreeBank, *in* 'Proc. of HLT 2002', San Diego, CA.

Kozlowski, Raymond, Kathleen F. McCoy & K. Vijay-Shanker (2002), Selectional restrictions in natural language sentence generation, *in* 'Proc. of Sixth World Multi Conference on Systemics, Cybernetics, and Informatics (SCI'02)', Orlando, FL.

Kunze, Claudia & Andreas Wagner (2001), Anwendungsperspektiven des GermaNet, eines lexikalisch-semantischen Netzes für das Deutsche, *in* B.Schröder, A.Storrer & I.Lemberg, eds, 'Probleme und Perspektiven computergestützter Lexikographie', Niemeyer, Tübingen, pp. 229–246.

Lakoff, George & Mark Johnson (1980), *Metaphors We Live By*, The University of Chicago Press, Chicago and London.

Lemnitzer, Lothar & Andreas Wagner (2004), Akquistion lexikalischen Wissens, *in* H.Lobin & L.Lemnitzer, eds, 'Texttechnologie. Perspektiven und Anwendungen', Stauffenburg, Tübingen, pp. 245–266.

Levin, Beth (1993), *English Verb Classes and Alternations*, The University of Chicago Press, Chicago and London.

Li, Hang (1996), A probabilistic disambiguation method based on psycholinguistic principles, *in* 'Proc. of 4th Workshop on Very Large Corpora (WVLC-4)', Copenhagen, pp. 141–154.

Li, Hang & Naoki Abe (1995), Generalizing case frames using a thesaurus and the MDL principle, *in* 'Proc. of RANLP-95', Tzigov Chark.

Li, Hang & Naoki Abe (1996), Learning dependencies between case frame slots, *in* 'Proc. of COLING'96', Copenhagen.

Li, Hang & Naoki Abe (1998), 'Generalizing case frames using a thesaurus and the MDL principle', *Computational Linguistics* **24**(2), 217–244.

Li, Ming & Paul Vitányi (1992), Inductive reasoning, *in* E. S.Ristad, ed., 'Language Computations', Vol. 17 of *Series in Discrete Mathematics and Theoretical Computer Science*, DIMACS, pp. 127–148.

Litkowski, Ken (2004), Senseval-3 task: Automatic labeling of semantic roles, *in* R.Mihalcea & P.Edmonds, eds, 'Senseval-3: Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text', Association for Computational Linguistics, Barcelona, pp. 9–12.

Manning, Christopher D. & Hinrich Schütze (1999), *Foundations of Statistical Natural Language Processing*, MIT Press, Cambridge, MA.

Marcus, Mitchell P., Beatrice Santorini & Mary Ann Marcinkiewicz (1993), 'Building a large annotated corpus of English: The Penn Treebank', *Computational Linguistics* **19**(1), 313–330.

McCarthy, Diana (1997), Word sense disambiguation for acquisition of selectional preferences, *in* 'ACL/EACL Workshop on Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications', Madrid, pp. 52–61.

McCarthy, Diana (2001), Lexical Acquisition at the Syntax-Semantics Interface: Diathesis Alternations, Subcategorization Frames and Selectional Preferences, PhD thesis, University of Sussex, Brighton.

McCarthy, Diana, John Carroll & Judita Preiss (2001), Disambiguating noun and verb senses using automatically acquired selectional preferences, *in* 'Proc. of SENSEVAL-2 Workshop at ACL/EACL'01', Toulouse.

McCawley, James D. (1968), The role of semantics in grammar, *in* E.Bach & R.Harms, eds, 'Universals in Linguistic Theory', Holt, Rinehart & Winston, New York, pp. 125–169.

Miller, George A., Claudia Leacock, Randee I. Tengi & Ross T. Bunker (1993), A semantic concordance, *in* 'Proc. of ARPA Workshop on Human Language Technology', San Fancisco, pp. 303–308.

Miller, George A. et al. (1990), Five papers on WordNet, CSL Report 43, Cognitive Science Laboratory, Princeton University.

Mitchell, Tom M. (1997), *Machine Learning*, McGraw-Hill, New York.

Osborne, Miles (1997), Minimisation, indifference and statistical language learning, *in* W.Daelemans, A.van den Bosch & A.Weijters, eds, 'Workshop Notes of the ECML/MLnet Workshop on Empirical Learning of Natural Language Processing Tasks', Prague, pp. 113–124.

Ostler, Nicholas & Sue Atkins (1992), Predictable meaning shift: Some linguistic properties of lexical implication rules, *in* J.Pustejovsky & S.Bergler, eds, 'Lexical Semantics and Commonsense Reasoning', Springer, New York, pp. 87–98.

Pereira, Fernando, Naftali Tishby & Lillian Lee (1993), Distributional clustering of English verbs, *in* 'Proc. of 31st Annual Meeting of the ACL', pp. 183–190.

Peters, Wim (1996), Corpus-based conceptual characterisation of verbal predicate structures, *in* 'Proc. of Computational Linguistics in the Netherlands', Antwerpen.

Peters, Wim & Ivonne Peters (2000), Lexicalized systematic polysemy in WordNet, *in* 'Proc. of LREC 2000', Athens.

Quinlan, J. Ross & Ronald L. Rivest (1989), 'Inferring Decision Trees Using the Minimum Description Length Principle', *Information and Computation* **80**, 227–248.

Rabiner, Lawrence R. (1989), A tutorial on hidden Markov models and selected applications in speech recognition, *in* 'Proc. of IEEE 77', pp. 257–286.

Resnik, Philip (1993), Selection and Information: A Class-Based Approach to Lexical Relationships, PhD thesis, University of Pennsylvania.

Resnik, Philip (1995*a*), Disambiguating noun groupings with respect to WordNet senses, *in* 'Proc. of 3rd Workshop on Very Large Corpora (WVLC-3)', Boston, MA.

Resnik, Philip (1995*b*), Using information content to evaluate semantic similarity in a taxonomy, *in* 'Proc. of 14th International Joint Conference on Artificial Intelligence (IJCAI)', Montréal.

Resnik, Philip (1997), Selectional preferences and sense disambiguation, *in* 'ACL SIGLEX Workshop on Tagging Text with Lexical Semantics: Why, What, and How?', Washington, D.C.

Resnik, Philip (1998), WordNet and class-based probabilities, *in* C.Fellbaum, ed., 'WordNet: An Electronic Lexical Database', MIT Press, Cambridge, MA, pp. 239–263.

Ribas, Francesc (1994), An experiment on learning appropriate selectional restrictions from a parsed corpus, *in* 'Proc. of COLING', Kyoto.

Ribas, Francesc (1995*a*), On Acquiring Appropriate Selectional Restrictions from Corpora Using a Semantic Taxonomy, PhD thesis, Universitat Politècnica de Catalunya.

Ribas, Francesc (1995*b*), On learning more appropriate selectional restrictions, *in* 'Proc. of 7th Confernece of EACL', Dublin, pp. 112–118.

Rissanen, Jorma (1989), *Stochastic Complexity in Statistical Inquiry*, World Scientific, New Jersey.

Rissanen, Jorma & Eric Sven Ristad (1992), Language acquisition in the MDL framework, *in* E.Ristad, ed., 'Language Computations', Vol. 17 of *Series in Discrete Mathematics and Theoretical Computer Science*, DIMACS, pp. 149–166.

Rooth, Mats (1998), Two-dimensional clusters in grammatical relations, *in* M.Rooth et al., eds, 'Inducing Lexicons with the EM Algorithm', Vol. 4 (3) of *AIMS*, Universität Stuttgart, pp. 7–24.

Rooth, Mats, Stefan Riezler, Detlef Prescher, Glenn Carroll & Franz Beil (1998), EM-based clustering for NLP applications, *in* M.Rooth et al., eds, 'Inducing Lexicons with the EM Algorithm', Vol. 4 (3) of *AIMS*, Universität Stuttgart, pp. 98–124.

Schulte im Walde, Sabine (1998*a*), Automatic semantic classification of verbs according to their alternation behaviour, Diplomarbeit, Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart.

Schulte im Walde, Sabine (1998*b*), Automatic semantic classification of verbs according to their alternation behaviour, *in* M.Rooth et al., eds, 'Inducing Lexicons with the EM Algorithm', Vol. 4 (3) of *AIMS*, Universität Stuttgart, pp. 55–74.

Schulte im Walde, Sabine (2003), GermaNet synsets as selectional preferences in semantic verb clustering, *in* 'Proc. of GermaNet-Workshop: Anwendungen des deutschen Wortnetzes in Theorie und Praxis', Tübingen.

Shannon, Claude E. (1948), 'A mathematical theory of communication', *Bell System Technical Journal* **27**, 379–423, 623–656.

Stetina, Jiri, Sadao Kurohashi & Makoto Nagao (1998), General word sense disambiguation method based on full sentential context, *in* 'Proc. of COLING/ACL'98 Workshop on Usage of WordNet for NLP', Montréal.

Ueberla, Joerg P. (1994), 'On using selectional restriction in language models for speech recognition'. **URL:** *http://arxiv.org/abs/cmp-lg/9408010*

Viegas, Evelyne (1999), Opening the world with active words and concept triggers, *in* E.Viegas, ed., 'Breadth and Depth of Semantic Lexicons', Kluwer Academic Publishers, Dordrecht, pp. 263–282.

Vossen, Piek, ed. (1999), *EuroWordNet Final Document*, EuroWordNet (LE2-4003, LE4-8328). Deliverable D032D033/2D014.

Wagner, Andreas (1995), Integration von Selektionsbeschränkungen in die LFG-Komponente von GTU, Diplomarbeit, Universität Koblenz-Landau.

Wagner, Andreas (2000), Enriching a lexical semantic net with selectional preferences by means of statistical corpus analysis, *in* 'Proc. of ECAI-2000 Workshop on Ontology Learning', Berlin, pp. 37–42.

Wagner, Andreas (2002), Learning thematic role relations for wordnets, *in* 'Proc. of ESSLI 2002 Workshop on Machine Learning Approaches in Computational Linguistics', Trento, pp. 99–113.

Wagner, Andreas (2003), Estimating frequency counts of concepts in multiple-inheritance hierarchies, *in* 'Proc. of GermaNet-Workshop: Anwendungen des deutschen Wortnetzes in Theorie und Praxis', Tübingen, pp. 69–78.

Wagner, Andreas & Claudia Kunze (1999), 'Integrating GermaNet into EuroWordNet, a multilingual lexical-semantic database', *Sprache und Datenverarbeitung. International Journal for Language Data Processing* **23**(2), 5–19.

Wagner, Andreas & Mattia Mastropietro (1996), Collecting and employing selectional restrictions, *in* 'Papers of the First Swiss-Estonian Student Workshop on Computational and Theoretical Linguistics', University of Tartu, pp. 76–85.

Wilks, Yorick (1986), An intelligent analyzer and understander of English, *in* B.Grosz, K.Spark Jones & B.Webber, eds, 'Readings in Natural Language Processing', Morgan Kaufmann Publishers, Los Altos, CA.