Polysemy

Adam Kilgarriff
D. Phil.
University of Sussex

December, 1992

Abstract

What does it mean to say a word has several meanings? On what grounds do lexicographers make their judgments about the number of meanings a word has? How do the senses a dictionary lists relate to the full range of ways a word might get used? How might NLP systems deal with multiple meanings? These are the questions the thesis addresses.

The 'Bank Model' of lexical ambiguity, in which polysemy is treated as homonymy, is shown to be flawed. Words do not in general have a finite number of discrete meanings which an ideal dictionary would list. A word has, in addition to its dictionary senses, an indefinite range of extended uses. The lexicographer describes only the uses which occur reasonably frequently and are not entirely predictable from the word's core meanings.

Polysemy is not a natural kind. It describes the crossroads between homonymy, collocation, analogy and alternation. (An alternation is a pattern in which a number of words share the same relationship between pairs of usage-types.) Any non-basic type of use for a word can be treated as belonging 349(or900.99.7ae0.3(uses99.3(349(o9.349(es.))]TJ191.286(8-999.349.7(88010Td(indefinite)aF3(a)-1499A

Contents

PREFACE

There are four kinds of thesis in cognitive science: formal, empirical, program-based

Chapter 1

Introduction

Many words have many meanings. The dictionary tells us so. But what does that mean? How many meanings does a word have, and what grounds do we or the lexicographers have for saying it is not more or less? And how is it that language-users can effortlessly comprehend and generate novel uses of words? What does that tell us about lexical structure? How might natural language processing computer systems deal with multiple meanings, or novel meanings?

These are the questions that motivate the thesis. The introductory chapter will spell out these questions in greater detail and show why we need a fuller understanding of polysemy; address the question, "What is polysemy?"; defend the kinds of methods used; take a tour of the thesis, sketching the methods and results of each stage of the research; and finally draw attention to the three principal claims the thesis makes.

1.1 Why is polysemy interesting?

1.1.1 All human knowledge ...

What is the structure of human knowledge? The question demands attention but is vast – far too vast to be directly researchable. The domain must be constrained: not 'knowledge' but some specific variety of knowledge. Some have taken knowledge of geometry, or geology, or arithmetic, or medical diagnosis: others have shifted focus from the knowledge itself to the words that are used to express it. Section 8.2 argues that lexical and general knowledge may share structure in some important respects.

So what is the structure of the human lexicon? But that is still a huge question. Words relate to other words in innumerable ways, and some corner of the whole must be selected for study. Again, the area of study must be reduced and focussed. Here, a methodological consideration comes to our assistance. A proven experimental technique is to hold as many factors as possible constant, in order that any observed variation can be attributed to a limited number of sources. In studying polysemy, we hold the form of the word constant, and then observe variations in meaning and distribution. Research into polysemy is one avenue for investigating the structure of lexical knowledge and hence, indirectly, the structure of human knowledge.

1.1.2 Wilks's problem

> LDOCE[1] is a full-sized dictionary containing over 55,000 entries, most of which contain multiple sense definitions. This level of real-world detail creates special problems of

[1] Longman Dictionary of Contemporary English

> scale for language processing systems. ... [E]ven a simple-seeming sentence like
>
> > There is a huge envelope of air around the surface of the earth.
>
> –considering only traditional content words (nouns, verbs, adjectives and adverbs)– represents a big space, because LDOCE contains 11 adverbial senses for *there*, 2 adjectival senses for *huge*, 14 nominal or verbal senses for *air*, 11 senses for *around*, 7 for *surface* and 12 for *earth*. Taken all together, and at the most crude level of analysis, this sentence is a a staggering 284,592-way ambiguous.
>
> This demonstration simply serves to show that a parser for text, accessing a realistic machine-readable language-resource like LDOCE, is faced with solving a large, and hard, problem. And not a problem *created* by large on-line dictionaries, rather a problem of *language* ... (Slator & Wilks, 1987, p 4–5)

Here is a second reason for studying polysemy. What is an NLP system to do when it goes to the dictionary to find the meaning of a word and finds several? Wilks's problem is one of the great obstacles lying in the way of wide-coverage natural-language computer systems. An account of polysemy is a prerequisite to removing that obstacle.

1.1.3 Creativity in language use

The fact that language users utter sentences that have never been uttered before has long been seen as one of the central facts linguistic theory must account for. It is a *sine qua non* of syntactic theory that it account for an indefinite number of possible sentences. But novelty in language use is not constrained to syntax. Language-users also use words in novel ways, the most dramatic beingf34.08010Tdmetaph(for)T46m--357.1266(b)-100f-h[(An)-1600ymTf34.805(einrhat)-126languagesynbfor.

Here, *bake*

thresholds of 'sufficiency' and 'insufficiency' will depend, for paper lexicography, on the size and target audience of the dictionary. In inheritance-based computational lexicons as described in this thesis, 'predictability' will be subsumed under a more general treatment of the inheritance structure of the lexicon.

Polsemy does not form any kind of 'natural kind'. It describes, rather, a crossroads. In one direction lies homonymy, in another – as in the 'highly predictable' case discussed above — metonymy. In others again, collocation and analogy. 'Collocation' describes those usage-types which only occur in the neighbourhood of one, or a small number of, other words, so the use is best described in a dictionary by giving the meaning of a multi-word unit and the question of whether there is a distinct sense for a particular word is side-stepped. 'Analogy' is used for those usage-types which are predictable, but the predictability stems from general knowledge and the situation of use rather than a rule which might be stated in the lexicon and considered a pattern of metonymy.

For each direction, there is no natural divide between polysemy and its neighbour. *Light*, of colour and of weight, may be considered homonymous or polysemous: *whisky*, used of the liquid ("a glass of whisky") or of a glass of it ("I'll have a whisky"), polysemy or metonymy: *light* in "travel light", polysemy or a collocation.[3] Polysemy is a concept at a crossroads and an investigation into it must be an investigation of the roads leading into and out of it.

1.3 A defence of the methodology

Before proceeding to a resumé of the contents of the thesis, some comments regarding the kinds of methods used are in order. Some would argue that the lexicon is an abstraction from the mental lexicon, which is best studied using psycholinguistic methods, so here we defend our approach against that attack.

1.3.1 A cognitive science perspective

> I believe, perhaps with undue pessimism, that the mind is too complicated to be seen clearly, or to be studied with advantage, from the perspective of a single discipline. The scientific understanding of cognition depends on a synthesis; [my research] is an attempt to bring together some of the ideas and methods of experimental psychology, linguistics and artificial intelligence. (Johnson-Laird, 1983, p xi)

We take the study of the lexicon to be intimately related to the study of the mind, and hold that these considerations also apply to it. For an understanding of the lexicon, the contributing disciplines are lexicography, psycholinguistics and theoretical, computational and corpus linguistics.

1.3.2 Conceptualism and realism

In some quarters, psycholinguistics is seen as having a privileged view of the lexicon, with other insights and information sources secondary. Thus, in building a lexical entry, Ilson & Mel'čuk (1989) say:

> Now, we believe that in contemporary English, BAKE is primarily a verb of cooking. This belief is based not on frequency but on psychological salience ...(p 336)

In a paper on 'WordNet, a lexical database organized on psycholinguistic principles', we find:

[3] Simple examples of analogy are hard to come by, for reasons discussed in section 7.4.

Pinker (1989, p 101) claims that speakers of English decompose verbs into such semantic subpredicates as CAUSE, GO BE and PATH

1.3.3 The benefits of formalism

dispelled. Different words are different because they have different spellings and sounds. There is no comparable fact of the matter for determining what makes a word sense different.

1.4.2 Breaking the Bank Model: Chapter 6

> "Read and read until you find something that everyone seems to accept that just can't possibly be true ..." (Anthony Robins, in conversation)

In addressing Wilks's problem, all researchers have, at least until the late 1980s, implicitly adopted the 'Bank Model', which we characterise as follows:

> A word like *bank* presents a very clear case of a word with more than one meaning. It can mean the side of a river, or an institution which looks after your money for you. For any usage of the word as a noun, either a money bank or a river bank is being referred to, and the word always refers to one or the other, not both. When English speakers encounter the word in a discourse, they know instantly and effortlessly which meaning of the word applies. This knowledge is an important part of human competence in a language, and an NLP system, likewise, needs to be able to choose.
>
> The word *bank* has been used because, in it, the issues are clear to see, making it a good pedagogical example.

classes of verbs is taken from Levin & Rappoport Hovav (1991) and formalised. Here, the different usage-types have different syntax. The analysis shows how the appropriate syntax-semantics mappings are inherited for each usage-type, and gives a concise, default-based account of the relations between syntactic complements and syntactic arguments. A further consideration was that the formalism in which an NLP system is written will impose constraints on the form the lexical entries should take. The lexicon entries described by the DATR theory inherited a DAG-like structure, making them directly usable by any NLP system working within an HPSG-like unification-based formalism (Shieber, 1986; Pollard & Sag, 1987).

1.5 Claims

The thesis makes three principal claims, one empirical, one theoretical, and one formal and computational. The first is that the Bank Model is fatally flawed. The second is that polysemy is a concept at a crossroads, which must be understood in terms of its relation to homonymy, alternations, collocations and analogy. The third is that many of the phenomena falling under the name of polysemy can be given a concise formal description in a manner that elucidates the relationship to alternations, metonymy and homonymy, and which is well-suited to computational applications. Two small parts of a lexicon that meets these goals are presented.

Chapter 2

Literature Review I: Word Sense Disambiguation

2.1 Introduction

A major theme of the thesis is how polysemy has been addressed within natural language processing. It has long been evident that there is a problem. Words often have several meanings. If an NLP system is to operate at all on the meanings of words, it will have to access the right meaning where there is a choice of several. This chapter will look at the history of the problem, reviewing the literature and sketching how findings from other disciplines shed light on the arena.

In the beginning, there was the direct approach. The problem was that words were ambiguous, so the solution was to devise procedures for disambiguating them. The concern was for showing what it was possible to do with computers, in the Artificial Intelligence mould. Section 2.2 covers this work.

But a dominant feature of the lexicon is its size. There is a 'lexical acquisition bottleneck'. Writing the procedures for disambiguating words was very time-consuming. For many, a more appealing technique was to extract information from an existing source: the dictionary. Machine-readable dictionary research is chronicled in section 2.3.

2.2 he Artificial Intelligence tradition

The work in this tradition has included the Bank Model amongst its theoretical presuppositions. For all this work,

1. The author gives no justification of how he chose the sample of words to be considered. The words are selected according to the researcher's ideas of what words were interestingly ambiguous.

2. The senses to be chosen between were arrived at by the investigator. All the authors make reference to the large numbers of senses to be found in dictionaries as a major source of motivation, yet no use is made of published dictionaries thereafter.

3. The possibility that a usage might fit more than one sense is not mentioned.

4. Only a very small number of words has been studied.

Chapter 6 considers these reservations in detail; for the time being, let us put them to one side.

referents, a preference semantics system is not equipped to make the leap. Wilks's contention was that such cases were relatively infrequent, and that the computational costs of resolving them resolvingtax012T0-ving

requi37(te-so-m]TJ5959(inists)]TJ92w)p-273.83918.4oEp(lesisITION)]-,I12T0-ving
4-5.6ving

> The construction of word experts requires patience, dedication, and finesse, and inherently involves far more intricate labor than ought to be expected of any person. (p 200)

Small's approach to parsing pays great regard to the idiosyncratic behaviour of words. The ubiquity of exceptions and oddities, and the extent to which much research has overlooked it, is a concern shared by the current study.

2.2.

and meaning of the sentence, *viz.*, what sorts of things are being referred to and into what sorts of roles do they fall. Hirst provides an architecture in which all these cues co-operate, and, as a part of the process of arriving at an interpretation of the input in a general-purpose

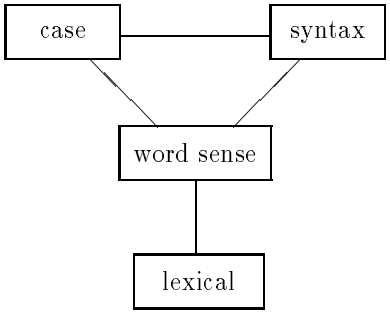


Figure 2.1: System architecture, Cottrell (1989), p 13.

the subject role, and between just one of the NPs and the AGENT case. In Cottrell's version of connectionism, any aspect of the interpretation of a sentence is to be represented by a unit firing. To allow for all the possibilities, the straightforward option is, therefore, to have a 'binding' unit, for every NP-and-syntactic-role pair, and for every NP-and-case-role pair. The unit would then fire only when that NP filled that role. The problem is, the number of these units will increase exponentially. Cottrell considers and suggests some partial solutions.

The model is clearly a complex and impressive system, and it offers some intriguing insights into, for example, different strategies for making prepositional phrase attachments. However Cottrell, like Small, is devising a system where one style of representation and processing is applied to syntactic, semantic, lexical, selection restriction and word association cues alike. His arguments for doing this are that, firstly, in our brains, all these things are going on in a similar medium, *viz.*, impulses along neurons, and secondly, the psycholinguistic evidence that the processing of all these cues is in parallel in humans. While these things are both true, they do not indicate that our ability to model or reproduce the behaviour will best be served by a connectionist system. An alternative strategy is 'divide and rule', whereby the syntactic, semantic and other aspects of the process are abstracted away, in order that they can receive specialist attention. Once these processes are better understood on their own, our chances of identifying and modelling the contribution they make to the overall problem will be much improved. Despite his claims to an interdisciplinary approach, Cottrell makes almost no reference to the linguistic literature, be it on syntax, formal semantics, verb valency or lexicology, although all of these areas have received a great deal of specialist attention which is relevant to the task. In due course, when what is specific and different about all these processes is sufficiently well understood, when, in Marr's terms (Marr, 1982), we have resolved what computations are being performed in the course of lexical disambiguation and according to what algorithms, then there will be a level of understanding which makes the question of how the algorithms are implemented in neurones a focussed and well-specified one. But until that day, a 'divide and rule' approach is But

and to arrive at a dictionary-based inheritance hierarchy. Disambiguation strategies are needed for the production of a lexical knowledge base from an MRD, and the information gatherered in that process can then also be made the content of lexical knowledge base entries, particularly since tasks the knowledge base might be used for, such as information retrieval, often involve sense disambiguation.

Klavans (1988) surveys the 1988 state of the art (from an IBM perspective). She emphasises the level of demand for theory-neutral computational lexicons and discusses how IBM's products, first UDICT and now CompLex, have fared. The demand is from the whole range of types of NLP application, from translation to query-answering and style-checking. She identifies two major problems, the polysemy problem and the mapping problem. The mapping problem is the problem of mapping the senses that one dictionary gives for a word onto the senses that another dictionary gives. Different dictionaries divide up the range of uses a word has in different ways. The problem is ubiquitous. The polysemy problem comes about when semantic features and selection restrictions are to be attached to dictionary word senses. The UDICT formalism only allowed one list of features for each word in a specific part of speech. Which features are to get onto that list: any feature that any usage of the word might reasonably have, or only those features that all usages of the word might be expected to have? Behind both major problems lurks the same ugly question: what are we to make of dictionary word senses? From Amsler's early work until now, it has sat stubbornly by.

In MRD research the problems and projects have tended to be practical, to do identifying and exploiting the structure of the published dictionary. It has not been a domain of competing theories but rather one where different research groups have developed various tools and techniques, or have achieved wide-coverage analyses, and further research has simply been able to use the results of what has been done before. What follows is a brief history of the projects completed and tools developed.

2.3.1 Amsler: *The Structure of the Merriam Webster Pocket Dictionary*

The first major work in the field was Amsler's thesis (Amsler, 1980). It was his goal to discover the conceptual hierarchy or hierarchies implicit in this particular MRD. His motivation was to provide a knowledge representation scheme, upon which inference could be performed, for a substantial part of the vocabulary of English. The method was computer-assisted, rather than automated or semi-automated. His extended example comprised the verb *move* and those verbs defined in terms of it. He worked from the assumption that all definitions complied with both the 'replaceability' criterion (that the definition should be able to replace the word without change of meaning) and a genus-and-differentiae structure. Definitions which did not fit these patterns were modified so that they did, in a pre-processing phase. Then, all the verbs with *move* as the genus terms in their definitions were collected. A search was made for all the different senses of *move* in the language, in the dictionary under scrutiny and elsewhere. Then, for each sense of *move*, the definitions in which *move* was being used in that sense were identified. This was one of the labour-intensive operations for which large numbers of paid disambiguators were brought in. He then looked through the definitions to determine what range of case relations were specified in the differentiae of the definitions. Then, with the range of cases known, the differentiae could be rewritten as sets of attribute-value pairs.

The work is important in that it showed that taxonomies could, in a systematic if not at that time automated process, be generated from dictionaries. It encounters and discusses the problems of loops in the dictionary; of definitions varying from the genus-and-differentiae structure; of hard-to-identify genus terms; and of the relation between case-relations and semantic primitives, thus setting the agenda for much subsequent research.

tory, IBM's Thomas J. Watson Research Centre, Illinois Institute of Technology, the Computing Research Laboratory at New Mexico State University, the Istituto di Linguistica Computazionale del CNR, Pisa, amongst others, has been setting about automating various of the processes that Amsler performed manually. Byrd, Calzolari, Chodorow, Klavans, Neff, & Rizk (1987), and other articles in the same special edition of Computational Linguistics describe progress so far, as does Boguraev & Briscoe (1989), a volume specific to research on LDOCE. Wilks, Fass, Guo, McDonald, Plate, & Slator (1989), in that volume, explicitly address the question of the potential of MRDs as resources for building lexical knowledge bases, and both the introduction to the book and that paper review the history of the enterprise.

The techniques include parsers and pattern matchers for dictionary definitions. They are aimed at identifying genus terms and their inverses (hypernym relations, or a term's 'children' in the taxonomy), lexical relations (Evens, 1988), and more specific sets of words such as active and stative verbs (Byrd et al., 1987). Different articles concentrate on different parts of speech (e. g. adjectives in Ahlswede (1985)), on definitions which vary from the genus-and-differentiae format (Vossen, Meijs, & den Broeder, 1989; Guthrie, Slator, Wilks, & Bruce, 1990), or on the exploitation of thesauri, corpora (see below) and further dictionaries (Calzolari, 1989).

Crucial to much of this work is the fact that lexicographers tend to use

subdivisions. The disambiguation process first parses the string to identify the verb and the noun which heads the NP. Each of these

Chapter 3

Literature Review II: Linguistics and Lexicography

3.1 Introduction

The last chapter surveys responses to Wilks's problem. But Wilks's problem was only one of several answers to 'Why study polysemy?' This chapter surveys work concerned with other answers.

As argued in the introduction, dictionaries are a treasure-house of information on polysemy. Perhaps their authors have direct answers to the questions of how, and with what rationale, lexicographers have deemed words polysemous. Section 3.2 considers the lexicographical literature.

The credibility of dictionaries rests on their authors' consideration

Zgusta warns him or her "not to be too impressed by the basic uncertainty concerning the nature of lexical meaning" (p 24) and, until recently, the methodology of researchers in linguistics has almost always been too contentious and the conclusions insufficiently specific to be of practical use. More recently, particularly since Hornby's Oxford Advanced Learners Dictionary (first published 1948), lexicography has been increasingly informed by the results of work on syntax in linguistics, but word meaning has not been a particularly fruitful area of linguistic research. So there the situation

the translation of *question* is *poser*, usually translated as *put* or *place*. The verb associated with *question* is not predictable from its meaning. The verb to be used must be stated in the lexicon for each language. The ECD has a lexical function, $\mathbf{Oper}_1$, for the relation from a noun to the semantically empty (or at least emptied) verb which it stands as direct object to, so we have in the ECD for English:

$\mathbf{Oper}_1$(QUESTION) = ASK

and in the ECD for French:

$\mathbf{Oper}_1$(QUESTION) = POSER

Mel'čuk and Polguère claim:

> [Lexical functions] and their combinations allow one to describe exhaustively and in a highly systematic way almost the whole of restricted lexical cooccurrence in natural languages. (p 272)

There are different varieties of lexical functions: some

3.3.1 Corpus lexicography: COBUILD

Collins COBUILD dictionary broke new ground in the 1980s through its commitment to basing dictionary entries on corpus evidence. The criterion for listing a word or phrase was, at least in part, that it occurred with sufficient frequency in the COBUILD corpus. The editors chronicled the difficulties faced and decisions made in the course of the project and published the account as a book, *Looking Up* (Sinclair, 1987). The chapters by Moon and Hanks provide a preview of many of the topics covered in this thesis, from the practical angle of 'should, or how should, this usage be represented in the dictionary'. They discuss briefly, *inter alia*, collocations, lexical fields, subcategorisation possibilities, metonymy, connotations, metaphoricity and literalness. If a word's usages fall into two classes with regard to any of these, there may well be a case for presenting alternative word senses. In COBUILD, as in most dictionaries, the presentation of an alternative word sense is not an all-or-none matter; there are a variety of strategies for showing different sorts of variation from a previously-defined sense (see also Chapter 5). The COBUILD project has been highly influential throughout lexicography, and the principle that dictionary entries should be based on corpus evidence is now widely accepted.

3.3.2 Atkins: semantic ID tags

Atkins, one of the originators of the COBUILD team, pursues the idea that "every distinct sense of a word is associated with a distinction in form" (Sinclair, 1987, p 89) in Atkins (1987, 1990). She extracts all the corpus citations for the word under scrutiny and proceeds to work through them, noting any patterns there may be in the relations between the meaning conveyed and the form of the utterance.

3.3.3 Tagging and parsing

An essential early stage of language processing, for a general purpose NLP system, is syntactic parsing. But syntactic parsing is an operation, not directly on the words in a text, but on the syntactic categories of those words. The process of assigning categories, or tags, to words is called 'tagging'. Many words (a 'word' is here simply a sequence of letters surrounded by blanks or punctuation) can belong to more than one category. Also, many of the word types in a corpus are not to be found in a large dictionary.[2] So tagging is not a trivial task.

The nature of the tagging and parsing tasks and the preprocessing required are discussed in various articles in Garside, Leech, & Sampson (1987), Meijs (1987) and Aarts & Meijs (1990).

3.3.4 Corpus as testbed

For the taggers, parsers and statistical models described above corpora are needed to test the theories as well as to build the models. This raises chicken-and-egg problems. While a corpus can be used to develop a parser, according to the best linguistic practice, it is no easy matter to determine whether it is doing the job properly, as there does not currently exist a large quantity of accurately parsed text against which the parser's performance can be compared. It is the goal of the Penn Treebank Project (Marcus, Santorini, & Magerman, 1990) to produce such a test corpus. The method is pure bootstrapping: as it is found to be far quicker for people to check a candidate parse than to come up with one themselves, a parser is used on a tagged corpus and trained people check and correct the computer's output. If certain kinds of mistake are consistently made, the parser or grammar can be improved. With the aid of a special purpose interface, the production of a corpus of parsed text or 'treebank' is well under way and soon there will be treebanks for testing parsers as well as parsers for building treebanks.

That is one use of a corpus. More and more, corpus evidence is required to support theory across linguistics. The researcher must ask, is the phenomenon found in the corpus? Do the circumstances in which it occurs tally with what the theory predicts? What proportion of the occurrences does the theory elucidate, and are the exceptions damaging to it? Whittemore, Ferrara, & Brunner (1990) test how successful different strategies for attaching prepositional phrases are, using a corpus of examples. Briscoe, Copestake, & Boguraev (1990, discussed below) trawl the corpus for all occurrences of a particular family of verbs, to test their theory.

Many varieties of linguistic hypothesis could be tested using a corpus. As yet few have been. Corpus-based studies provide an opportunity for the empirical examination of many linguistic questions, and it is to be expected that a wider and wider range of issues will be broached using corpus-based methods.

3.3.5 Corpus-based sense disambiguation

As corpus analysis tools are developed, so the corpus becomes a source which can be exploited for sense disambiguation. McRoy's work (section 2.2.6) used some corpus-based strategies. Hearst (1991) is highly corpus-based. The program is restricted to homograph disambiguation: it does not attempt to discriminate between all the senses given in a dictionary, but only between those that are markedly different. The program works from a list of clues for disambiguation, which might be called 'potential semantic ID tags'. These include whether the word is capitalised, whether it modifies another item, and whether it is found in a PP headed by one of *in, on* and *of*. In the learning phase, the program is fed a substantial number of corpus citations with the target word disambiguated. By seeing which potential clues apply, with what frequencies, to citations of the different senses, the program is able to determine which items of information are salient for disambiguation. Once the relative importance of the clues has been established, new inputs can be fed to the program which will calculate and compare the evidence for each of the possible senses for the word. Hearst tests her system, and finds the results "comparable to, or better than earlier efforts using MRD's and large corpora" (p 19) such as Lesk's, Guthrie et al.'s and Veronis and Ide's. She briefly notes the difficulties of making quantitative assessments and comparisons.

3.3.6 Collocations

A collocation is a group of two or more words which are to be found in proximity to each other significantly more often than one would predict, given the frequency of occurrence of each word

co-occurring words for near-synonyms such as *strong* and *powerful*, but while the presence of a word in both lists is evidence of a type of usage that both words share, the absence of a word from one of the lists is only the weakest of evidence that the use of the two words differ in respect of the third. A sharper tool for investigating how, or whether, two words differ in respect of their frequency of co-occurrence with a third is the t-score, which looks at the difference in frequency of occurrence of, e.g., *strong tea* and *powerful tea* and counts how many standard deviations away from each other they would be if they were both part of the same population. It can serve to highlight, for lexicographers, contexts for which words of similar meaning have different distributions and can thus lead to greater definition in those parts of the dictionary entry where a word is set apart from its semantic neighbours.

With the automatic language generation goal in sight, Smajda & McKeown (1990) have developed a program, Xtract, for extracting collocations from a corpus and representing them in a form convenient for re-use in a language generation program. This use requires rather fuller accounts of the collocations than does lexicography. The generation system will need to know whether the words in the collocation must, or may, occur in a given order, with a given array of other words falling between the words in the collocation. For collocations involving verbs, it will need to know what sort of argument slots the phrase leaves, where in the string they are to be realised, and what selection restrictions apply. Xtract does not merely identify collocations; it automatically builds lexical entries for a phrasal lexicon. Zernik & Dyer (1987) describes a similar project, though not in the context of corpora.

3.3.8 Hindle: combining parsing and statistics

The mutual information and t-test statistics, when applied to unparsed text, look only at fixed two-word strings. Results of interest relied on there being a grammatical relation of interest -—usually of modifier to head— between a word and its neighbour. But for most grammatical relations, we do not expect to find the two related words next to each other. The text must be parsed for word pairs standing in particular grammatical relations to each other to be identified. Then the statistical tools can be applied to populations of grammatically-related word pairs, and the lexicographer can gain a far fuller picture of the kinds of settings the word is to be found in.

A paper pursuing this line of research is Hindle (1990). His premise is that semantically similar

as its object[4]

(c) Take the minimum of the two mutual information scores [5]

(d) Repeating stages a–c but with 'object' replaced by 'subject';

(e)

They identify six different possible relationships between transitive and intransitive forms of a verb. Sometimes more than one of these applies to the same verb ("John is baking."). The pattern or patterns the verb participates in depend largely though not entirely on the core meaning of the verb. Levin (1991) argues:

> As the distinctive behavior of verbs with respect to diathesis alternations arises from

to be interpreted should be assigned to the pragmatic interpretation of the word in context. He

have not merely this sense of *kill*, but also the metaphor structure, available in its

Generative Lexicon enterprise to avoid redundancy by specifying generative mechanisms by which non-basic senses are implicit in the lexicon without being explicitly listed.

Part of what the usage-types have in common is that, in both cases, the phrase subcategorised for denotes an event. Verb phrases in -ing typically denote events, as do noun phrases headed by *event*, *party*, *flight* and the like, so *enjoy* constructions with these elements need no further explanation. However, *the paper* ordinarily denotes a physical object (or a non-count mass). When appearing as object of *enjoy*, it must be 'coerced'. The entry for the relevant sense of *paper* in the dictionary makes reference to reading (and writing), and introspection confirms that the default event associated with a paper is reading (or possibly writing) it. Composition with a verb such as *enjoy* foregrounds this information and makes available to the dictionary reader, or NLP system with a Generative Lexicon, a 'reading(/writing) the paper', event-designating sense of *the paper*.

Pustejovsky calls the purpose typically associated with a noun its 'telic role', where TELIC is one of the slots in the qualia structure. Telic roles are explored further in an experiment described in the Briscoe et al. paper. They investigate how widely the analysis given for *enjoy the paper* can be applied. They take the group of verbs, including *enjoy* and also *prefer*, *finish*, *start* and *miss*, which take both NP and infinitival or progressive VP complements. The hypothesis is that when the verb takes an NP and the head noun is not a word for an activity, it will generally be a word for an object which is either designed for or otherwise intimately associated with a particular activity, and thus has a well-defined telic role. In 'unmarked' usages, where there is no event-type supplied by the context, the original sentence needs reinterpretation if the clash between event-demanding verb and non-event-denoting object is to be resolved. The solution is to interpret the sentence as an expanded version in which the present participle of the default activity verb is inserted between the original verb and object. Hence, since *cigarette* has the telic role of being smoked and *film* of being watched, "finished his cigarette" becomes "finished smoking his cigarette" and "missed the film" becomes "missed watching the film".

They test the theory by trawling the LOB corpus for examples of the verbs under scrutiny, and then examining the examples with non-activity noun phrases as complements to see if there is a telic role associated with the head noun, and if so, whether this was what was being enjoyed, preferred or finished. Their results support the hypothesis. There was a readily available telic role for nouns appearing in this direct object slot in most cases, and the exceptions did seem to be 'marked' cases in which the context supplied an event-type.

According to Pustejovsky, coercion and corresponding logical metonymy do not apply only to this case. Adjectives such as *fast* are also taken to modify the telic role. If the sense of *fast* in which it modifies motion is taken as primary, then, in "a fast car" it is the driving that is fast (in this primary sense) and in "a fast typist" it is the typing (which is the telic role of the word if not of the typist his/herself!). Pustejovsky describes three other roles, alongside the telic one, and considers various constructions alongside the *enjoy* one which, he claims, coerce nouns into one or other of these roles. However the combination of telic role and verb of the *enjoy* class is the only construction where the case for logical metonymy has been worked out in detail and empirically tested. For *fast* and other examples presented in Pustejovsky (1989) it is not clear that either the roles which would need to (a)Tj9.1199212Td[6uot

phenomena. Thirdly, semantic composition must be conceived sufficiently broadly so that it can incorporate the process of determining, from the lexical entry for a noun, which role (i.e. telic, or other) is required in a particular context.

The Generative Lexicon brings together several themes in contemporary work on lexical semantics, computational linguistics and NLP. It is still very much a set of ideas and small scale implementations in an early stage of development, and many of the ideas are still to be worked out, but does present an intriguing list of desiderata for NLP lexicons.

Chapter 4

Theory I: What is a Word Sense?

4.1 Introduction

In learning a language, the learner arrives at the meaning of a word through identifying what there is in common to the role the word plays in all the various discourses he or she has heard it in. We follow situation semantics (Barwise & Perry, 1983) in taking the meaning or meanings of a word to be an abstraction from the role(s) it systematically plays in discourses.

Pollard & Sag (1987) present the view as follows:

> According to situation semantics, the world is made up of such things as *individuals* (like Jon Barwise or the moon), *properties* (such as being a cookie or being a donkey), *relations* (such as seeing or kicking) as well as *situations*. Roughly, situations are limited parts of the world which consist of individuals having (or not having) properties, or being (or not being) in relations. An example of a situation is the particular event of Carl Pollard eating a certain orange in Office D-2 at Ventura Hall, Stanford University, at 9:42 p.m. PST, December 2, 1986. Individuals, relations, properties, and situations are real, but different groups of organisms are *attuned* to different ones in accordance with the exigencies of their ecology; as it is sometimes put, different communities of creatures "tear the universe apart along different seams."
>
> What does this have to do with meaning? According to situation semantics, meaning arises from *constraints* that hold between different kinds of situations. For example, any situation that has smoke in it is part of a situation that has fire in it. We say that smoke *means* fire; any organism that is attuned to this constant can pick up from a smokey situation the information that there is fire. So it is with language, but in that case the constraints involved are not *natural* ones; rather, they are *conventional* linguistic constraints that can be exploited by the people that are attuned to them, as when an English speaker acquires from an utterance of "Here is a cookie" the information that there is a cookie. Linguistic meaning, then, is a relation that holds between types of utterance situations and the types of things in the world that utterances describe. This view is called the *relational theory of meaning*. (pp 4–5)

This 999.csco00.3(fles)Tj13.68012999.4(do)-10004(fireth)Tj23.51990Td(.24]TJ90.7199y)0.7(b)-1000(eopl24Tf17.5

and then reinterpret it in realist terms:

> the *signifiant* is a certain type of utterance situation, namely one where "cookie" is uttered; the *signifié* is a certain property of things in the world, namely the property of being a cookie. And the sign itself is not a psychological association but the real-world linguistic-meaning relation (a constraint) between the *signifiant* and *signifié*. On this view, a natural language such as English is not a shared mental system but rather a type of linguistic-meaning situation in which certain conventional constraints are observed. The English "cookie" sign, then, is a subtype of English linguistic-meaning situation: the type of situation where "cookie" is used to mean cookie. (p 5)

A word's meaning is learnt when a speaker uses the word in accordance with the constraints, so his or her utterances of the word bear the same relations to situations as other speakers'. For lexicography, the task is to identify the constraints holding between a word and the situations where it is used. This involves looking at populations of usages, and seeing what, in addition to the occurrence of the word in question, they have in common.

We shall follow Lyons (1977) in using 'denotation' to refer to the part of the world corresponding to a word, whereas *referent* will be used for the entity a word relates to in a particular utterance.

Formal semantics generally operates with a model in which denotations are assumed as the entities to which compositional rules apply. For example the referent of a definite noun phrase such as "the cookie" is a member of the set denoted by the head noun (*cookie*). The denotation is a given, from which the analysis proceeds. However for our purposes the denotation cannot be a given. Do different senses have different denotations? The question merely reiterates the puzzle of the nature of polysemy. While the meaning/denotation distinction is of central importance for model theoretic semantics, it is of little relevance to the investigation of polysemy.

4.2 Nunberg's Referring Functions

Nunberg (1978) shares the concern that a view of meanings as denotations leaves much unsaid. He analyses the view that a lexical entry is a specification of the things and classes of things that a word can normally be used to refer to. He takes examples such as:

> (1) Hearst bought a newspaper.

and points to the ambiguity between it being the copy or the company that Hearst has bought. If a lexical entry specified a class of things, then since the class of newspaper copies and the class of newspaper companies are so clearly disjoint, either a disjunctif9920a4(from)-1Tis

interpretation in both cases, a theoretical framework in which the two are treated differently is unsatisfactory. Pragmatics is an area of language-understanding requiring general-purpose inference mechanisms, whereas the lexicon is often seen as supporting only simple look-up procedures. Nunberg shows that both pointing and a choice of a word can serve the purpose of determining reference, and that in both cases, inference mechanisms are required.

The path from the word *newspaper*, or the object pointed at, to the referent can be simple or complex. It may involve any aspects of the verbal or non-verbal context, or of mutual knowledge of speaker and hearer. The task of specifying the lexical entry is doomed, unless it is accepted that there is always a further job to be done, for which inference is required, of establishing what
.

The answer lies in the facts of how words are used. The extent to which speakers of a language use RFs other than those which are standardly used with a particular word is an entirely empirical question. The lexicographer's task is to list all tolerably common usage-types. If he or she has done that, then usages which do not fit the lexicographer's usage-types will be relatively rare. We might say that the RF from a word to its familiar usage-types is a well-trodden path. A speaker will be more readily understood if he or she gives the hearer directions only along well-trodden paths (except in cases where short-cuts are easy to see). Thus speakers will, most of the time, exploit already-familiar routes. So, alongside the core meaning of the word in the lexical entry, there needs to be a catalogue of the RFs which are found to be used with the word.

Nunberg's arguments first suggested that lexicons listing large numbers of senses for words were doomed to failure as they could not list all usage-types. Indeed, lexical semantics cannot be separated from the pragmatics of reference and the lexicon alone will never list all usage-types, but, directly or via a referring function and an inference mechanism, it does need to represent all those used with significant frequency.

4.3 MacWhinney's Competition Model

MacWhinney (1989) presents a model of lexical acquisition which gives an account of the dynamics of RFs, sense extension, and similar. His 'Competition Model' explores the process of categorisation. It aims to show "how competition provides a way of understanding the semantic ranges of words" and "how words force each other to take on various polysemic and extended meanings" (p 195). The model is a connectionist one. The task it models is one of lexical choice: it is trained to associate different situations in which a choice is to be made, modelled as sets of input features, with different 'word' outputs. In the testing phase, it is fed a set of features and selects a word as output. It is a 'competition' model because the words that are the output options 'compete' to be the output associated with a new input.

MacWhinney contrasts his model with two other models of concepts and categorisation which have been centre-stage in recent times. Firstly, the 'Classical Theory', according to which there are necessary and sufficient conditions for an entity to fall under a concept. To know the concept is to know those conditions in order that entities can be classified as falling within it or not falling within it.

Second, 'Prototype theory'. Psychological evidence, notably that produced in a series of

concepts. The 'Competition Model' is so called because alternative concepts are 'in competition' for as yet unclassified objects which might or might not fall under them, so the coverage of a concept is determined by its interaction with other concepts in the neighbourhood. This spatial metaphor is pervasive in MacWhinney's work. In his words:

> The notion of semantic topography is a useful way of understanding the ways in which words compete for meanings. This topography makes distinctions not just between words, but also within words. (p 213)
>
> Major polysemic entries can be seen as corresponding to the valleys of large rivers in this topography, whereas major and minor polysemes correspond to increasingly smaller tributaries. Determining the exact reading for a given word is like tracing a a stream back to its source. Some of the decisions are easy and can be made on the basis of the words in the sentence. Other decisions require rich situational information or prior discourse cues. (p 215)

The metaphor takes usages of words in referring expressions as its starting-point and has limitations elsewhere, but does provide a useful vocabulary for describing the domain. The process MacWhinney models is that of selecting a word that an entity or situation 'falls under': metaphorically, determining for an input grid reference, which valley is most accessible. The word can then be used to denote the entity or situation in the future. The competition model suggests there are areas in semantic space, the valleys, committed to particular words, but there are other areas —high ground,

& McClelland, 1986). The investigator identifies what seem to be the salient dimensions of the semantic space under consideration, the 'features'. The connectionist network is then 'trained' by presenting it with pairings of a word and a bundle of feature-values. A network can be trained in relation to a number of different words. When the training phase is complete, if the network is presented a bundle of feature-values as input, its output will correspond to the word which gives the 'best match'. Features could, in principle, relate to both linguistic context and non-linguistic settings in which the word was used: they might cover semantic and pragmatic aspects of the situation, as well as syntactic, phonetic and collocational ones. Thus the Competition Model offers an account of how words are learnt, by analogy with how the network is trained. It offers an account of how the word meaning is represented. And the network, given incomplete

words. The lexeme is the entity the dictionary entry is about, which, in the case of the verb *drive*, includes *drive, drives, driving, drove* and *driven*. Lyons (p 22) states two preconditions for homonymy. Firstly, only lexemes of the same part of speech can be homonyms, and secondly, to be homonyms, all forms must be shared. Thus forms with the same spelling but different parts of speech are ruled out from being forms of homonymous lexemes. Identical forms having different parts of speech clearly are related in meaning in some cases (*rust* as verb and noun) and not, in others (*can* in "You can do it!" and in "tin can"). The effect of Lyons's stipulation is simply to make any such concerns a separate topic to homonymy and polysemy. The thesis takes a similar line, and does not consider at length issues about the relationship between verbal and nominal *rust* or *can* (the matter falls under derivational morphology) or the disambiguation problems they present to NLP systems (which are largely solved —see section 3.3.3). We do not adopt Lyons's terminology: while 'lexeme' is an important construct for morphology, the distinctions between the different morphological forms of a word are not central to a discussion of polysemy: we talk of 'words' not 'lexemes'.

An etymological criterion is simply that a sense pair corresponds to two different words if at an earlier point in the history of the language the two variants' precursors were different in sound or spelling. Thus *mouth* (of a person) and *mouth* (of a cave) share histories, and

variation in register or dialect, but that would not provide a reason for lexicographers and linguists to say there were two different senses of *dream* even if the dreams described in one register were distinct from those described in the other. The case for seeing the verb as ambiguous would at least need supporting evidence from, for example, other verbs from the same domain —*imagine, fantasise, visualise*

This is clearly not falsified by the failure of the Central Reserve Bank to guarantee certain river banks. However,

Yesterday we sold all the red jumpers in the shop.

is falsified if all the pillar-box red, but not all the brick red jumpers were sold.

Against the test stands the question of what we would ordinarily expect a quantified expression to range over. Presumably the banks guaranteed by the Central Reserve Bank in the example do not include the Central Reserve Bank itself, yet that does not effectively falsify the statement once we know that central banks are special kinds of banks which need not be included in the scope of quantification for the statement. Likewise, foreign owned banks might be excluded. The scope of quantification is commonly fixed (inasmuch as it is determinate at all) pragmatically, by the context and the kind of thing being said, rather than by any purely-conceived meaning of the word. Thus the interpretation of quantified statements always involves an assessment of the scope of the quantifier. The widest scope is rarely what is called for. Yet the test assumes the widest scope possible, and that the boundaries of 'the widest scope possible' defines the point at which vagueness turns into ambiguity. A test sentence is no longer simply one in which the target word is in the scope of quantification, as many such sentences are not falsified where it might seem they should be, as in the Central Reserve Bank failing to guarantee itself. The test sentence needs to be one in which the quantification is truly universal, but these items are rarer, and our intuitive assessments of them correspondingly weaker.

Conjunction

The conjunction test works on the premise that a single occurrence of an ambiguous word can only get one reading. It sets up sentences where the single occurrence participates in two sets of relations, as specified in conjoined clauses. The sentence must use the same occurrence of the target word, yet without requiring it to have the same referent in both clauses. Depending on the syntax of the expression to be tested, the effect can be achieved with conjunction alone, or in association with anaphoric *one*, *so did* or ellipsis. The question then is, is the reading where the one clause requires one use of the target word, and the other, the other, acceptable? If it is the case is one of vagueness, not ambiguity. Where the word is ambiguous, the expression has a kind of oddity known as 'zeugma' and often used in witticisms such as:

? She came out in spots and a bath chair.

The zeugma is evidence of the ambiguity of *come out*. One using our old familiar *bank* is:

?

4.4.3 General considerations regarding ambiguity tests

The tests have all been presented with the aid of an unproblematical example of ambiguity and an unproblematical example of vagueness. This was done in order to demonstrate what the test was and what the two contrasting outcomes were. However, in those cases the tests would never be necessary. What we want of a test is that it is consistent with our intuitions, where our intuitions are clear, and that it resolves the question, where our intuitions are unclear. The conjunction and quantification tests fare well in meeting the consistency condition. But do the tests help where intuitions are unclear? There does not seem to be any evidence that they do. Two problems

The conclusion is not surprising. A word has more than one meaning where the process of abstracting the contribution it makes to sentence-meaning produces two clusters of contribution-types. Where there are two entirely distinct clusters, a word is ambiguous, but where it is vague or unspecified, there is a single cluster which spreads out along some dimension. There is no *a priori* reason to expect to find any clear distinction between the two types of cases.

4.5 he ideal lexicographer and the essential word sense

There are no decisive diagnostic tests for identifying word senses, yet there is clearly an important role for word senses in NLP and lexicography. Some words have a number of distinct, conventionalised patterns of use and both human dictionary-users and NLP systems need them specified. Dictionaries provides a huge store of data about them. Lexicographers clearly have intuitions or strategies which enable them to do a fair job of identifying senses. We may ask how the lexicographers go about it.

An idealisation of lexicographic practice provides a working definition of a word sense, as follows. We assume that for each word, the lexicographer

1. gathers a corpus of citations for the word;
2. divides the citations up into clusters, so that, as far as possible, all the members of each cluster have more in common with any other member of that cluster, than with any member of any other cluster;
3. for each cluster, works out what it is that makes its members belong together;
4. takes these conclusions and codes them in the highly constrained language of a dictionary definition.

The process is an idealisation of what actually happens in dictionary-making, displayed to expose 'the central core of the lexicographer's art, the analysis of the [citation] material collected' (Krishnamurthy, 1987, p 75). Now that extensive corpora are available to lexicographers (at least in English and some other languages), lexicography is moving towards the idealisation. It focuses on a process of clustering usages, performed by a lexicographer. The lexicographer was probably not explicitly aware of the criteria according to which he or she clustered at the time, and stage 3 is a fallible *post hoc* attempt to make the criteria explicit. Yet it is those criteria which determine the senses that eventually appear in the dictionary. They are a result of that process. But they are a result at several removes, and with each of these removes comes the possibility of confusion or error.

The idealisation is of use for our search for the nature of word senses. We should like to know what they are, and where one ends and the next begins. 'No entity without identity' runs Quine's test, and without identity conditions for word senses the concept remains hazardously ill-defined. The idealisation points us towards the criteria the lexicographer was using for his or her clustering, because, however quirky they may have been, they are the data that the published form of the dictionary is attempting to communicate. They answer, as well as anything can, the Quinean test. The identity test for a word sense in a particular dictionary is that two usages of the word belong to it if and only if the lexicographer would have put them in the same cluster.

4.6 Sufficient frequencyl35() 47.04020 d(tns)fficien

judgements. These judgements are a major source of evidence regarding the structure of the lexicon, but what more general account of word senses may they be subsumed under? The answer is the SFIP criterion, introduced in section 1.2.

Every dictionary has a length limit, and yet research on the number of word types found in corpora (Walker & Amsler, 1986) does not indicate any upper bound on the number of words which are candidates for inclusion. Not all will fit, so some criteria are required for selection. An obvious criterion is frequency. Words which are used only very rarely can be omitted from any dictionary but one on the scale of the OED. The same consideration applies to word senses. A word sense must be of sufficient frequency, for it to earn its place in a dictionary.

The previous section indicates another criterion. Where usages fall into a tight cluster, a sense is defined to cover the cluster. Conversely, to be exemplars of a distinct sense, usages must fall outside clusters for already-acknowledged senses. The cluster is, at a first pass, a set of usages which are similar to each other, so the criterion for a distinct sense is that it must represent a cluster of usages which are related to each other but sufficiently dissimilar to any existing senses. In the case of brick red and pillar-box red, there are clearly not two distinct, dissimilar clusters of usages. There are merely two small subsets of the usages of *red* which are indistinguishable from the complete set for *red* (as a colour word) in terms of syntax, collocational possibilities, illocutionary force and other factors, but which denote specific parts of the range of colour which *red* denotes. The only way these subsets can be iden.3(Td(for)die)-13000(oi)-12000.bh tpecififyng ahesi senoteaion.

There

Chapter 5

Dictionary Study I: An Analysis of Word Sense Distinctions

5.1 Introduction

This chapter reports on an empirical study in which the distinctions between word senses were investigated. The range of mechanisms employed by the lexicographers for describing alternative usage-types was examined, and a classification scheme developed. A sample of words was selected, and for each of the words, if it had more than one sense listed in the particular dictionary chosen for the study, then the distinction(s) between those two (or more) senses became part of the population of word sense distinctions. The study succeeded in finding some distinct types of sense distinctions, but also indicated the heterogeneity of the distinctions, with the majority defying any simple classification.

Preliminaries

5.1.1 Choosing a dictionary

The dictionary chosen for the studies was

study.

5.1.2 Limiting the domain

To constrain and focus the topic, several kinds of meaning distinction will be excluded from the population of distinctions under scrutiny (inevitably, the boundaries will sometimes be hard to place). These are:

1. Part-of-speech ambiguit/e:

2. Numbered senses within an entry.

   This is the basic level at which meanings are explicitly distinguished.

3. Subdivisions of numbered entries marked **a**, **b**, etc.

   The rationale for using subdivisions rather than main divisions is not clear. A necessary but not sufficient condition for their use is that the senses distinguished in this way are more closely related to each other than they are to other senses.

4. Bracketed optional part.

   One sense is given by including the bracketed material, and another by excluding it.

   Examples:

   marquetry (the art of making) a type of pattern in wood ... .

   martini (a glass of) an alcoholic drink ...

   mazurka (a piece of quick lively music for) a Polish dance.

5. (fig.) in front of an example.

   According to the User's Guide,

   > Some words are used in an imaginative or "figurative" way, to suggest a meaning that is not the literal meaning but has some similarities with it. If a word is often used like this, the examples will include a figurative use, and this is shown by the note (fig.) (page F36)

   Example:

   materialize 1 (fig.) *I'd arranged to meet him at seven, but he never materialized.*

   There will always be a continuum between figurative uses of a word and distinct senses which have their origins in figurative usage, since at any point in time many word usage patterns will be in flux between extremes of originality and conventionality. Where a lexicographer draws the line between the two types of cases will always be somewhat arbitrary.

6. The main definition contains a disjunction.

   Examples:

   masked 2 by or for people wearing masks: *a masked ball*

   melody 1 a song or tune: *a haunting melody*

   A disjunction of this kind is not necessarily signaled by 'or'. It could be signaled by 'and', or by a comma. The semi-colon is used for the similar398Td'.Tf854(a)1000.01(14000.73ograp10000(and)-100th

6.

The square-bracketed code means the word can be a count or a non-count noun. The non-count, 'marriage is a good institution' reading of marriage might be taken as a distinct sense. The alternatives indicated by the disjunctive grammar code are often both illustrated in the examples.

8. Brackets with a disjunction in them.

   As well as indicating an optional extra meaning, brackets sometimes give selection restriction-like information on how a word is normally used, indicating what sort of an entity a noun or adjective is used to describe, or for a verb, what its subject or object is likely to be. When these brackets contain a disjunction of dissimilar kinds of entity, it could be said that two senses are being conveyed.

   Example: **meek** (of people or behaviour) ...

   The use of 'or' in brackets of this type is often an indicator of a range of possibilities rather than an indicator of disjunction, and the only disjunction it is regularly used to convey is the 'of people or behaviour' one used here.

5.2 Pilot study

The purpose of the study was to establish what, if any, were the commonly occurring patterns of distinctions between word senses. Approximately 1% of the dictionary, twelve of LDOCE's 1227 pages, were examined. The study was manual.

The sample contained 427 full entries. An initial analysis is given below.

Full entries	427	
of which - capitalised		20
- compound/phrasal/hyphenated		37
leaving a base population of	370	
2 separate entries in same word class	6	
Entries divided into numbered meanings[1]	110	
of which - 2 meanings		75
- 3 meanings		17
- 4 meanings		10
- 5 or more meanings		8
nouns		63
verbs		21
adjectives		25
adverbs		1
Total of numbered word sense distinctions[2]	187	
Numbered entries subdivided using letters	10	
Brackets give another sense (as in 4 above)	23	
of which - 'cause to'		4
- 'too'		3
Another word sense given as "fig" example	10	

Notes:

[1] Excluding numbered meanings where a phrase, idiom or collocation is given, also ignoring subdivisions of numbered senses by **a, b** etc.

[2] On the assumption that, where there are more than two senses, the number of sense distinctions worth considering is the minimum possible, i.e., one less than the number of senses.

5.2.1 Clarifying dictionary definitions

shortcoming: it provides an instance, but does not provide any indication of how much variation away from the example presented

5.2.3 Results and interpretation

(From analysis of distinctions between numbered word senses only)

	Nouns	Verbs	Adjectives	Totals
Genmets	9	6	3	

the expressive

Chapter 6

Dictionary Study II: Do Dictionary Senses Match Corpus Usages?

6.1 Introduction

This chapter describes an empirical study investigating how widely the Bank Model holds. A sample of words was selected using an effectively random procedure. For each, a set of citations was gathered from the LOB corpus. For each citation, an assessment was made as to whether it fitted one, none, or several of the senses given for the word in a desk dictionary. It was established that, for most words, some of their usages could not be classified into one and only one of the senses the dictionary gave. Usages were often indeterminate between senses, and senses were often insufficiently clearly identified for it to be possible to classify with any confidence. The study indicates that the Bank Model is limited in its range of relevance, and hence that NLP needs alternative models of how the usages of a word relate to the kinds of senses a dictionary provides for it.

For the purposes of the study, a usage is specified by a corpus citation of about eighty characters with the word under scrutiny in the middle. Thus the only available 'context' for a usage is verbal and is given in the (on average) seven preceding and seven following words. This notion of 'usage' is clearly a very limited one.

6.2 Identifying the sample

The dictionary used was, once again, LDOCE. The source of usages was the LOB corpus. The sample of words to be investigated was arrived at in the following way. Very common words were excluded because they tend to have very large numbers of senses and to present complex and difficult cases. For this study, simpler cases were to be examined. Low frequency words were excluded because it would not be possible to see any patterns emerge unless there were a reasonable number of usages to be examined. A range meeting these constraints was 26-29. So the initial sample was chosen by taking all those words which had between 26 and 29 occurrences in the first half of the LOB corpus. Half of these, a sample of 154, were taken for further analysis.[1]

[1] The reason half the corpus, and half the sample, were taken was so that if a model were to be built on the basis of the study of the sample in relation to the usages in the first half of the corpus, then the untouched halves would provide an environment for testing the model.

From this set the following were removed; prepositions and adverbs (there would only have been two of these, not enough to make any general comments), proper names, adjectives relating to countries ('Dutch', 'Greek'), titles ('Earl', 'Congress'), and non-base forms of words ('cutting', 'created', 'directors') or forms which were base but where a non-base form occurs much more frequently than the base form. The size of the filtered sample was 83.

The rationale of the last points was that dictionaries generally list only base forms, so it was

6.3 wo worked examples

There follow two accounts of how usages were found not to fit one and only one of the senses. First:

> **image***n*
> **1** [C (**of**)] a picture formed in the mind: *She had a clear image of how she would look in twenty years time.*
> **2** [C] a picture formed of an object in front of a mirror or LENS, such as the picture formed on the film inside a camera or one's REFLECTION in a mirror
> **3** [C] the general opinion about a person, organization, etc., that has been formed or intentionally created in people's minds: *The government will have to improve its image if it wants to win the next general election.* | *The company tries to project an image of being innovative and progressive.*
> **4** [(*the*) S (**of**)] a copy: *He's the (very) image of his father.*
> **5** [*the*+S+**of**] a phrase giving an idea of something in a poetical form, esp. a METAPHOR or SIMILE
> **6** *old use* likeness; form: *According to the bible, man was made in the image of God.*
> —*see also* MIRROR IMAGE, SPITTING IMAGE

The citations included the following.

> of the Garonne, which becomes an unforgettable image. This is a very individual film, mannered,

Here the 'image' is an image/1, a picture in someone's mind (probably the author's; possibly the whole cinema-audience's).

yet this citation seems to cut across that supposed correlation[4].) Sense 3 can be set to one side on the basis that we have no evidence that the film was made by the Garonne tourist board (or similar). But we would rather not have to make a choice between senses 1 and 2. The usage makes reference to both the projected image/2 and the images/1 that the projected images/2 caused in people's minds, and to make a choice would be to reject half the story.

Second:

exercise *n*
1

that the dictionary provides. It is not a problem for a person, be they a native or a non-native speaker of the language, if more than one dictionary sense is able to resolve their uncertainty about what a word is contributing to the meaning of a sentence where it puzzles them. A user need only read an entry up to the point where their puzzlement is resolved[5]. Even if a word sense coming later in the entry would have fitted the usage equally well or better, such a user has no need of that further information. Hence it is not surprising, and no criticism of lexicographers, if usages can often not be assigned one and only one word sense. There is no practical or theoretical reason why they should be.

6.4 Results

Of a sample size of 83 words, 14 had just one sense for each part of speech they featured in, leaving 69 for which there was potentially a choice to be made. For 60 of these 69 words, there was at least one usage which could not with any confidence be classified into one rather than an other of the senses. Thus the sense selection task presented in the experiment could sometimes not be resolved to a single sense for 87% of words where the possibility arose.[6] Appendix A. contains specifications of which words fell in which categories.

Sceptics and advocates of the Bank Model may argue that another researcher would have succeeded in classifying all or nearly all usages. Appendix B presents definitions and concordance lines which, in this experimenter's opinion, could not be satisfactorily classified as one and only one sense. The challenge for the sceptic is then to identify how each of the usages in Appendix B should be treated.

The usages which could not be assigned one and only one sense could have been classified according to the reasons they defied classification. This was not done in this case because the range of explanations was arrived at only in the course of conducting the experiment. For a further experiment it would be a worthwhile exercise.

6.5 Observations

The first point to make is that the exercise was, much of the time, hard. In the Bank Model people select senses instantly and effortlessly. For the sample of words chosen here, the experimenter was frequently toiling laboriously.

The task was hard in the cases where more than one of the dictionary senses was near the usage in the citation. The dictionary provides only a set of clues to the nature of the senses that the lexicographer was intending to discriminate. Identifying the divisions that the lexicographer saw in the conceptual space of usages of a word is a matter of reconstruction, and the citations are essential to the task, so as experimenter I was working at clarifying the sense distinctions throughout the process of classifying usages. It was not possible to work with an unchanging conception of the distinctions. Each time a new citation neither clearly fitted one and only one sense, nor replicated a pattern already seen, a re-evaluation of the sense distinctions for the word was inevitable.

The issues encountered varied greatly from word to word.

6.6 Conclusion

The sort of variation in meaning illustrated by 'bank' is far from typical of the kinds of variation to be found in the dictionary. Where a word has more than one dictionary sense it is relatively uncommon that each of a small sample of usages of the word can be classified into one and only one of the senses. There are many occasions where a word is used in a way which is indeterminate between alternative senses of the word, or the sense distinctions are not made sufficiently clearly in the dictionary to permit sagespvaethevthedoonarythe seno.3(d(6-14999.6ord12999.7(1ord12rnam14962(3lh.7(sensest0

Chapter 7

Theory II: A Four-Way Analysis

7.1 Introduction

This chapter brings together the arguments of Chapter 4 and the evidence of Chapters 5 and 6 to complete the account of the nature of polysemy. First we develop the discussion in the introduction of the crossroads nature of polysemy, then illustrate it with a worked example: *at breakfast*. Of the four signposted destinations, two, homonymy and collocations, have a familiar role in dictionaries. Alternations were discovered in the dictionary in the guise of 'bracketed optional parts' in Chapter 5 and are treated extensively in the next two chapters. That leaves analogy, and the next section considers why and where this strategy is needed. Closely related to analogy are those mainstays of so many discussions of innovation in word use, metaphor and metonymy. Section 7.5 places them within the analysis of this thesis. Both analogy and alternation depend on the notion of a semantic field, and section 7.6 provides a brief account of the main difficulties relating to that notion. Finally, we consider the prospects for the 'four ways' of the analysis coming together, so arbitrary choices as to how a usage-type is treated could be avoided.

7.2 he four ways

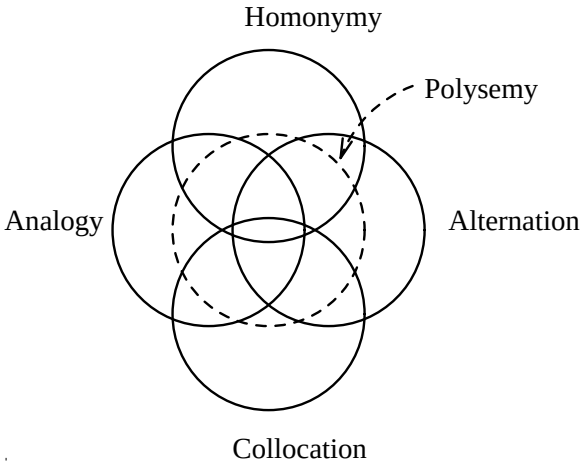


Figure 7.1: Polysemy and its neighbours

is to match the appropriate left hand side(s) for a usage in an input text. The task here is not to describe the form or content of the right hand side, nor to explain how the information might be gathered (though both tasks are closely related to the task at hand when a usage-type is implicitly rather than explicitly represented). The task at hand is simply expressing all the usage-types, that is, making available all the possible left hand sides. We shall only be concerned with any associated information in as much as it is necessary for identifying what usage-type we are expressing. The four ways of treating different cases of *prima facie* polysemy are, then, four ways of representing usage-types.

They are:

1. Homonymy. This is the straightforward case. Usage-types are expressed through being listed. They are represented in the lexicon as distinct one-word entries.

2. Alternation. A system of rules indicates how a non-basic usage-types may be inferred from a basic one.

3. Collocations. For a usage-type which co-occurs only with a limited range of words, all the collocations are usagsn(2.000.4(usage(sorting(all(4600.3(ar9.3(lexicon0(indi/44000.9J559.76020756[(co-o)-99

Word	Freq.	*at*	*to*	*for*	*before*	*after*	others	Total	Rel. freq.(%)
breakfast	54	2	3	2	1	5	1	14	26
lunch	66	1	8	5	2	9	4	29	44
dinner	99	11	2	12	4	9	1	39	39

The position adopted throughout the thesis is that usage-types

low enough. If metonymy is a device used in order to refer, then, the more obvious the RF (see section 4.2) from word to referent, the more effective it will be. For successful reference, it is essential that the hearer chooses the RF intended by the speaker. So the speaker must predict the hearer's choice, and the hearer, retrospectively, the speaker's. To the extent that the basis for prediction is a familiar pattern in the language, the RF will be represented in the lexicon as an alternation: metonymic usages will be interpreted by the same mechanisms as alternations. To the extent that the basis for the prediction is to be found in the particular situation which supplies a one-off RF, the prediction can only be made on the basis of that context and the relevant usage-type can only be made available by analogy.

Martin's work on metaphor is discussed in section 3.6.1. He specifically — and wisely — addresses only conventional metaphor. Conventional metaphors follow conventions, so are fairly predictable. Martin treats conventional metaphor as a set of rules which maps out a search space of possible metaphoric readings. Since there are rules to map out that search space, there are rules which render the readings predictable.

Many prototypically metaphoric uses, like that of *pig*, are sufficiently institutionalised to be listed in paper dictionaries. LDOCE gives,

> **pig** ...**2** *infml derog a* an unpleasant person, esp. who eats too much, behaves in an offensive way, or refuses to consider others

There is something paradoxical about this. The metaphoric use of *pig* is one in which a person is conceived of as an animal. But the corresponding sense is listed in the dictionary with genus term 'person', so if a dictionary user retrieves that sense from the dictionary, there is not so much as a mention of the animal so there is no 'conceiving of one thing as another' to be done. We might posit that the mental lexicon, like the dictionary, stores common usage-types so does not need to reinterpret them on each occasion of use. Conventional metaphor is often in this trap. To the extent that it is conventionalised, it is no longer, in the most interesting, pragmatic sense, metaphor.

In Romeo's

> It is the East, and Juliet is the sun

the use of *East* and *sun* are clearly metaphorical in that, typically, when first encountered, they provokes a new understanding of love, Juliet, dawn and the east. The usage-types will not be listed in the lexicon, or available by alternation. The role of the lexicon is to represent patterns of use of words. But the power of the novel, striking metaphor is dependent on it defying the normal pattern of use for the word. It requires the hearer to go beyond the familiar uses and patterns which are coded in the lexicon, and to bring their imaginative and intellectual capacities to bear. To the extent that a metaphor is novel and striking, it is no longer, in any interesting sense, represented in the lexicon. It is still theoretically acounted for by analogy, but usage-types expressed '[l(patterns)-147a4000.3(the)-[(exp1.7598Td[(I(new)Tja000(and)-he)-[(exp1acouically)-ots)-14000315000.4

some purposes, be put aside. But in general, similarity is a matter of degree, and is a complex notion on which it is unlikely that people will have strong or consistent intuitions.

Homonymy and polysemy are major issues for any similarity measure. *Bishop*, *knight* and *rook* all feature in a 'chess' semantic field, but, outside the chess domain, they are not in the same field. If we could first resolve matters of polysemy and homonymy, that would assist the identification of semantic fields. But semantic fields are needed for the resolution of polysemy and homonymy.

The topic of semantic classification is a large and important one. It has been looked at from many angles, some statistical, some seeking taxonomies of concepts (Sparck Jones, 1986; Berlin, 1978; Dahlgren, 1988; Beckwith et al., 1991, see also sections 3.3.8 and 8.2). The reasons given at the beginning of the thesis for studying polysemy could readily be revised to give reasons for studying semantic fields. The two topics are closely related. Both are central to lexical structure. The point here is simply that all the strategies for analysing unfamiliar word-uses depend on semantic classification, but there remain many unanswered questions as to how this might best be done.

7.7 Prospects for a unified approach

In the best of all NLP systems, there would be seamless joins between the four strategies for representing putative polysemy. The whole would be described in one all-encompassing language and the question, "Which strategy is most appropriate here?", would make no more sense, in relation to a case like *at breakfast*, than "Is orange red or yellow?"

The next chapter points the way to such a seamless join between homonymy and alternations. A smooth join between analogy and collocation methods can also be sketched: an analogy for *finger*, in "keep one's finger in", is identified by searching a lexicon for collocations with a similar form to "keep one's finger in" but with one of the words switched for some other word in the same semantic field. An analogy is treated as a modified version of a collocation. An algorithm for interpreting 'analogy' usage-types would require data about semantic fields or near neighbours and a lexicon containing many collocations. It would then seek the best fit, involving as near neighbours as possible, between a collocation in the lexicon and the input string.

What, then, are the chances for bringing these two pairs of approaches under a common roof?

The prognosis on this question is mixed. Alternation provides a rule-based approach to interpreting unfamiliar usages. Analogy deals with rule-defying cases, so cannot use a yes/no criterion for whether a use-pattern is available for a word since the answer by the rules would always be 'no'. It must use measures of closeness and goodnalogy.

Chapter 8

Formal Lexicography I: Noun Alternations

8.1 Introduction

This chapter takes those cases of polysemy that are susceptible to treatment as alternations, and presents a formal method for concisely capturing the generalisations and, potentially, making them available for exploitation for lexicography and NLP.

To recap: alternations, or 'regular polysemy' (Apresjan, 1974) occur where two or more words each have two senses, and all the words exhibit the same relationship between the two senses. An example, taken direct from LDOCE, is:

> **gin** (a glass of) a colourless strong alcoholic drink ...
>
> **martini** (a glass of) an alcoholic drink ...[1]

In each case, there are two senses referred to, one with the 'bracketed optional part' included in the definition and the other with it omitted, and the relation between the two is the same in both cases.

Recent work on lexical description has stressed the need for the structure of a lexical knowledge base (LKB) to reflect the structure of the lexicon and for the LKB to incorporate productive rules, so the rulebound ways in which words may be used are captured without the lexicon needing to list all options for all words (Boguraev & Levin, 1990; Gazdar, 1987; Pustejovsky, 1991a). The generalisations regarding regular polysemy should be expressed in the LKB, and the formalism in which the LKB is written should be such that, once the generalisation is stated, the specific cases follow as consequences of the inference rules of the formalism.

As 'lexicalism', the doctrine that the bulk of the information about the behaviour of words should be located in the lexicon, has become popular amongst unification grammarians, so formalisms for expressing lexical information have been developed. Some part of the syntax, semantics and morphology of most words is shared with that of many others, so the first desideratum for any such formalism is to provide a mechanism for stating information just once in such a way that it is defined for large numbers of words. Inheritance networks serve this purpose. The next requirement is that exceptions and subregularities can be expressed. It must be possible to describe concisely the situation where a word or class of words are members of some superclass, and share the regular characteristics of the superclass in most respects, but have different values for some feature or cluster of features. Several lexical representation formalisms addressing

[1] As the LDOCE entry for *glass* notes, a receptacle need not be made of glass to be a glass.

these desiderata have been proposed (DATR: (Evans & Gazdar, 1989a, 1989b, 1990), Russell, Ballim, Carroll, & Armstrong-Warwick (1991), Copestake (1991)). While the generalisations to be formalised are better understood for morphology and syntax, the theoretical gains, of capturing generalisations and eliminating redundancy, and the practical benefits, in terms of lexicon acquisition and maintenance, apply also to regular polysemy.

The work described here will take the DATR formalism and use it to represent a collection of facts and generalisations regarding polysemy. This chapter and the next use DATR, and a brief description will be given below, but the thesis neither presupposes a knowledge of it, nor gives a formal description. Evidence regarding regular polysemy will be introduced, in stages, with the proposed DATR account of the evidence worked through at each stage. The sense that results from excluding the bracketed part, or which is listed first in the dictionary, or which is the only one listed, will be deemed the 'primary' sense, with others 'secondary'.

8.2 axonomies of words and of their denotations

In the fragment presented here, facts about both the word and its denotation are accessed through the same node in the inheritance network. Thus a query regarding the syntax of the word *beech*, and a query asking what type of thing a beech tree is, will both be made at the same node. It might be argued that this is to confuse two different kinds of information. The position taken here is that there is much to be gained from holding the two types of information together, and to keep them separate is to forgo opportunities for expressing and exploiting generalisations, and to force a wide range of arbitrary decisions and duplication. The position is related to the central tenet of cognitive linguistics (see section 3.5) that linguistic meaning must be studied in the context of the overall cognitive system, though the claim here is clearly of much narrower scope.

As described in Chapter 2, Amsler (1980) and many others have shown the dictionary embodies a taxonomy. The taxonomy is primarily a taxonomy of denotations, and unearthing a rudimentary structure for human general knowledge, for use in AI knowledge representation, was Amsler's goal. The non-linguistic knowledge in a monolingual English dictionary is stated in English and the labels for the nodes in the taxonomy are English words, so in the course of expressing non-linguistic facts about beeches and trees, the dictionary provides *tree* as the genus term for *beech*, thus alluding to the potential for inheritance between the words as a side-effect. The example sentences given for *tree* (sense 1) in LDOCE include *to climb a tree; to plant a tree; to cut down/chop down a tree*, and a dictionary user would be correct in interpreting these as collocations in which *beech* can substitute for *tree*. Whether this is at all a consequence of linguistic as opposed to encyclopedic facts is possibly an unanswerable question, but it is not a matter which need concern the dictionary-user since the encyclopedic cargo and the linguistic vessel are both making equivalent journeys.

Collocational information is one kind of linguistic information which is, to a substantial degree, predictable from word meaning. The subject of this chapter, regular polysemy, is another. *Martini* participates in the glass-of/drink alternation because martini is a drink, and if we discover a new drink called foobaz we know we can order a 'count' foobaz as well as drink a lot of 'mass' foobaz. Alternations, by definition, apply to classes of words. The classes are formed according to the words' denotations, and words will generally participate in the same alternations if their meanings or denotations are similar. Taxonomies, whether in biology, dictionaries, or AI knowledge representation group similar things together, and non-linguistic taxonomies will often identify the classes of words to which alternations apply.

LDOCE commonly uses the taxonomy defined by genus terms to express alternations applicable to both the genus term and its subordinates. Thus in

> **daffodil** a very common bell-shaped pale yellow flower ...

dahlia a big brightly-coloured garden flower ...

flower 1 the part of a plant, often beautiful and coloured, that produces seeds or fruit ... **2** a plant that is grown for the beauty of this part ...

LDOCE encodes implicitly that both *daffodil* and *dahlia* can be used in both the first and second senses of *flower*. The dictionary depends on the reader's ability to spot that senses 1 and 2 are intimately related so, all else being equal, they will be inherited as a pair. One or other of the senses for *dahlia* and *daffodil* has failed to meet the SFIP criterion (section 4.6), in that it is insufficiently frequent and unpredictable to be separately listed in the dictionary, but with an intelligent reading of the dictionary's taxonomy, both senses can be seen as implicitly present.

Building general-purpose taxonomies for NLP, thesauri such as Roget's, AI knowledge representation or other purposes is a difficult and lengthy process, in which there are various different sources of evidence to be brought to bear in attempting to reach principled decisions regarding the overall shape of the taxonomy (see. *e.g.*, Dahlgren (1988)). The lexicons of one or more languages inevitably play a large role in determining what the structure of the taxonomy should be. It is often impossible to say whether the words used in such a process are simply acting as carriers of non-linguistic, conceptual information, or are colouring the conceptual information with the particularities of a word, language or family of languages (Quine, 1960). The process of building a taxonomy of words is, thus, inextricable from the process of building a taxonomy of things. Evidence from all sources is needed for a single taxonomy. So a further argument for attaching lexical information to a general-purpose taxonomy is that both emerge from substantially the same evidence, and dividing the evidence between two structures will amount to weakening the empirical support for both.

A related point is that, where an alternation applies to a class of words in a general-purpose taxonomy, the situations where the alternation applies are likely to be motivated as well as described. If the lexical taxonomy only contained information about words, then the fact that all drink-words participate in the glass-of/drink alternation could be described, but could not be related to the 'explanation' that this was because they all denoted drinks, since the relation between the denotation of *martini* and that of *gin* would not be represented. A framework in which lexical and general knowledge are held together offers a better chance of relating the classes of words to which alternations apply, to classes identified for independent reasons, which in turn improves the chances of predicting what alternations apply to a word from a rudimentary knowledge of its use.

An argument against holding lexical and general knowledge in the same taxonomy is that the lexicon contains idiosyncrasies in a way in which the non-linguistic world does not. Thus *oats* and *wheat* have similar denotations, occupying similar locations in a taxonomy of the natural world, but in the lexicon, one is singular while the other is plural (Palmer, 1976, p 119). For simple inheritance systems this may well present difficulties, but that merely indicates the inadequacy of such simple systems. DATR is designed around the need to express exceptions as well as regularities. It was devised with lexical representation particularly in mind, and has not yet been used for representing general knowledge beyond toy examples, so its potential in this direction is as yet unexplored, but it makes clear that the exception-ridden nature of the lexicon is not a reason for regarding it as outside the scope of knowledge representation languages.

There are, then, several arguments for using a single taxonomy for words and for things. First, it will avoid 'territorial' issues: "Is this fact lexical or general?". Second, it will avoid the need to duplicate information and structure between two parallel taxonomies. Third, the sources of evidence for building the two in any case overlap. Fourth, a single taxonomy will facilitate the comparison of classes identified on lexical and non-lexical grounds, giving more scope for predictive power.

is adopted.

Once the

specifies that we inherit the value from `<path2>` at `Node2`.

As well as local inheritance, there is global inheritance. DATR stores a 'global context' node and path, and where a DATR sentence specifies global inheritance, then information is inherited from the global context. Double-quotes are used. Thus:

```
Node1:<path3>   == "<path4>".
```

specifies that the value for `path3` at `Node1` is inherited from the value of `path4` at the global context node (and the global context path is changed to `path4`). When a query is made to a DATR theory, global contexts are initialised to their values in that initial query, and 'quoted paths' such as `"<path4>"` are often used to say, "go back to the node the query was initially made

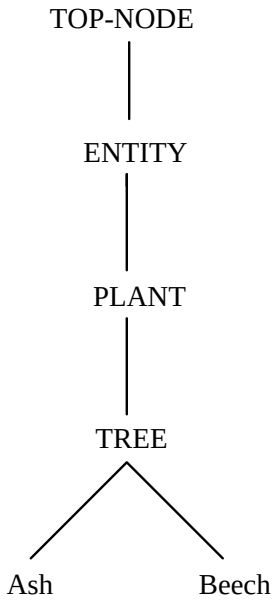


Figure 8.1: A simple taxonomy

This is shown as a taxonomy in Fig. 8.1.

To that basic structure, we wish to add a generalisation about 'wood' senses. Once we have established that *ash* is being used in its wood sense, we wish to treat the word as we would *teak* and *mahogany*. We need to distinguish secondary senses from primary ones in such a way that the paths for accessing information about them are different. We do this by prefixing the path with `alt` (for alternation). There might be several alternations, so we identify the alternation by the path element following `alt`, the 'label', for which we shall use the genus term of the alternate sense, here `wood`. Let us also add some flesh to the bare bones of the taxonomy, and state some genus terms, `word` values (i.e. the word associated with the node), and `collocates`, words commonly found as near neighbours of the target word, at various low-level nodes. The next version of the theory, to be explained below, is:[3]

```
TOP-NODE: <collocates>     == .
ENTITY: <>                 == TOP-NODE.
PLANT: <>                  == ENTITY.
TREE: <>                   == PLANT
    <collocates>           == plant grow chop-down PLANT
    <genus>                == tree
    <alt wood>             == WOOD:<>.
Ash: <>                    == TREE
    <word>                 == ash
    <alt wood collocates>  == black TREE.
Beech: <>                  == TREE
    <word>                 == beech.
```

[3] In accounts of DATR published to date, sequences are enclosed in round brackets. However the brackets are redundant in that they can be omitted without ambiguity, and future definitions of the language will not include them, so they have not been included here.

and this illustrates there is a wood/tree as well as a tree/wood alternation. Lexicographers have not used the same formula for *teak* as for *ash* and *beech*, and this corresponds to the fact that the 'wood' sense is the more salient for the former, the 'tree' sense for the latter. To represent the two patterns as the same would be to throw away a principled distinction made by the lexicographer.[4]

In this fragment, all alternations are represented as directional links and the relation between the two alternations is not expressed. The following code adds *teak*, and the wood/tree alternation, to the fragment.

```
WOOD:<>                 == ENTITY
    <genus>             == wood
    <collocates>        == table desk ENTITY
    <alt tree>          ==
```

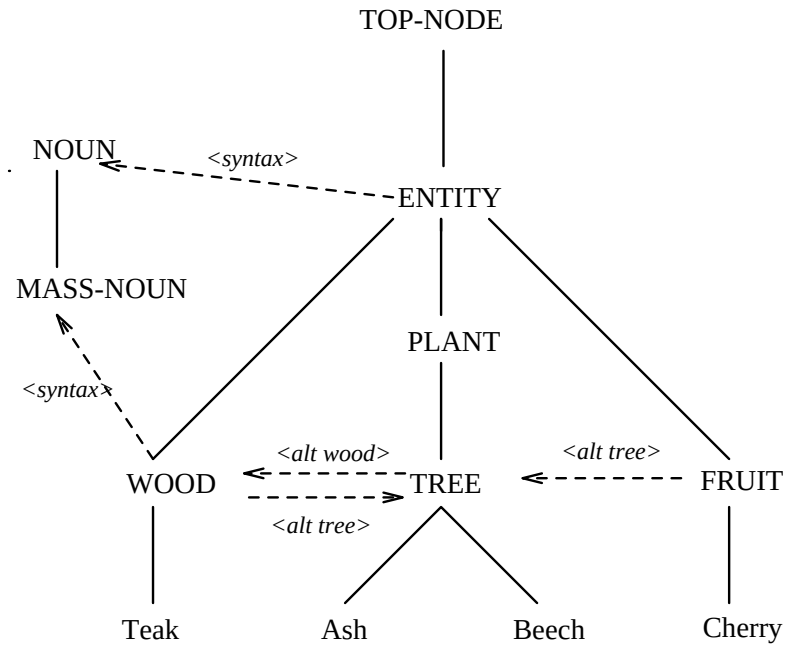


Figure 8.3: Taxonomy showing transitive alternations

```
Cherry:<alt tree alt wood genus>            = wood.
Cherry:<alt tree syntax count>              = yes.
Cherry:<alt tree alt wood syntax count>     = no.
```

There may be any number of `<alt x>`

Here we have a fruit/plant alternation, a more general variant of the fruit/tree pattern. If a `Strawberry` node which inherits by default from `FRUIT` is added to the theory, then, if we query the node with a path commencing `<alt tree>` or `<alt tree alt wood>`, the inheritance follows exactly the same course as for `Cherry`. If there are values for syntax, collocations or anything else which `Cherry` will pick up from higher up the `TREE` or `WOOD` parts of the taxonomy, then `Strawberry` will also pick them up. The theory above, when supplemented with

```
Strawberry:<> == FRUIT
    <word>    == strawberry.
```

has theorems

```
Strawberry:<alt tree genus>                    = tree.
Strawberry:<alt tree syntax count>             = yes.
Strawberry:<alt tree alt wood genus>           = wood.
Strawberry:<alt tree alt wood syntax count>    = no.
```

One direct approach to this problem would be to insert a `TREE-FRUIT` node between nodes for tree-growing fruit and `FRUIT`, and for that to be the point at which paths starting `<alt tree>` were redirected to `TREE`. However, the motivation for such a node is weak. 'Tree-growing fruit' is not used in the dictionary as a genus term,

```
TREE:<altlist>                  == wood PLANT.
FRUIT:<altlist>                 == "<grows-on>" ENTITY.
WOOD:<altlist>                  == tree WOOD.
FRUIT :<grows-on>               == plant.
Cherry:<grows-on>               == tree.
```

all take us to the same location in the inheritance network. It cannot be assumed that the 'looping' path will never be used. A recent innovation of the fish-marketing industry is to reconstitute the flesh of the fish into fish-shapes (and coat it in breadcrumbs for a quick-fried family meal). When the parent asks the child "how many fish do you want?" there is clearly one alternation from animal to food in evidence, and another which re-converts the mass, 'food' sense into a countable, 'fish' sense, yet the characteristics of the breadcrumbed version accessed through `Fish:<alt food alt fish>` are clearly not all the same as those of the original, and we might expect to find specifications such as

```
Fish: <alt food alt fish manufacturer> == Bird's_Eye.
```

even though the default case is that `Fish:<alt food alt fish` *path*`>` inherits from `Fish: <` *path*`>` in a loop-like structure. So apparently looping paths may occasionally give rise to interesting theorems, though usually they will not.[7]

8.7 Polysemy and homonymy in DATR

As discussed in Chapters 1 and 4, 'homonymy' and 'polysemy' are both useful concepts for the description of the lexicon of a language, although it is impossible to cleanly distinguish them. How might this state of affairs be modelled in DATR? A simple model of a DATR lexicon is one in which there is a node for each word or lexeme. Then, a natural treatment for homonymy is to carry information, and make queries, about different words at different nodes. But if polysemy often cannot be distinguished from homonymy, should polysemy not be treated similarly, with distinct polysemous senses each having their own DATR node?

The idea has some appeal, but it would mean that information about a word sense was never defined in the theory unless a node for the sense had explicitly been added. The fragment presented has said nothing explicitly about the 'tree' or 'wood' senses of *cherry* yet it represents facts about their syntax, genus terms and collocations. The senses are predictable and follow from generalisations and should not need explicit mention. It seems likely DATR theories will need to continue working with the homonymy/polysemy distinction, with homonyms treated as distinct nodes, and polysemous senses as sets of node-path pairs with a distinct path prefix. This offers the kind of flexibility required. To turn a sense treated as polysemous into one treated as homonymous is trivial. If we add

```
Cherry/tree: <>       == Cherry:<alt tree>
    <word>            == cherry.
```

to the theory, we now retrieve the same values for theorems of the form
`Cherry2: <`*path*`>`
as for ones of the form
`Cherry: <alt tree` *path*`>`.[8]
All the polysemous senses described in the fragment above can be treated as homonyms, each with a distinct node, in this way.

If distinct nodes were wanted for a pre-defined range of regularly-polysemous or homonymous senses, they could be generated automatically. (The node names would then be automatically generated so we might expect them to take a form `Word1, Word2` etc. rather than the mnemonic `Cherry2`. Since node

was being used in a previously-unfamiliar sense, so the theory would develop as it 'learnt' from its input.

8.9 Summary

A lexical knowledge base needs inference mechanisms, and a structure which reflects the structure of the lexical knowledge it conveys. DATR is a default inheritance formalism designed specifically for lexical representation. Regular polysemy is one level of structure in the lexicon, about which a desk dictionary provides an ample supply of facts. In this chapter we have examined and formalised the regular polysemy of a very small fragment of English. We have been able to exploit a number of generalisations about the domain to make the theory compact and productive. The formalisation both presents a theory of the operation of regular polysemy in one corner of the lexicon, and is a model for how regular polysemy might be used to structure a lexical knowledge base.

Chapter 9

Formal Lexicography II: Verb Alternations

9.1 Introduction

The previous chapter presented a formalisation of the polysemous behaviour of some nouns. The domain chosen there displayed no very interesting syntactic behaviour and did not address how lexical entries might meet the constraints imposed by a grammar formalism which makes parsing and semantic interpretation possible. This chapter formalises the alternations apparent in a part of the verb lexicon, taking these constraints into consideration.

The kind of phenomenon to be captured is the relation between *bake* in "John is baking the cake", "John is baking", and "The cake is baking".[1] Here are three different usage-types for the verb. It should not be necessary to introduce three different primitives into the lexicon. The alternations are regular and are shared with other verbs, so should be described at some general node in the taxonomy, and inherited.

One part of the exercise is to make the relations between the

Simply stating that a verb such as *bake* can be transitive or intransitive, and then giving example sentences, is not an adequate treatment in a learner's dictionary. How, AKL ask, is the learner to discover that "The girl washed" means the same as "The girl washed herself" rather than "The girl washed something".

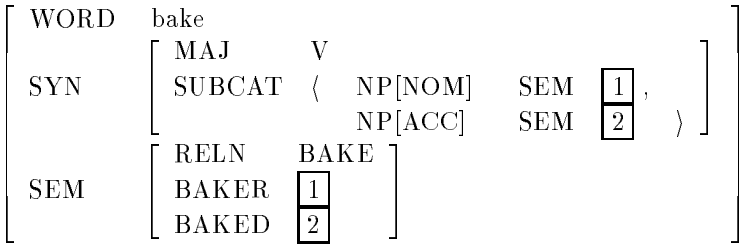


Figure 9.1: HPSG-like AVM for *bake*

Figure 9.2: AVM for NP [NOM]

Figure 9.3: AVM for NP [ACC]

Angle brackets are used for lists. List members are separated by commas and the components comprising a list member, here NP[NOM] and SEM 1 in the first case and NP[ACC] and SEM 2 in the second, are to be understood conjunctively, each being a partial description of the list member. To spell the lists out as feature structures, we adopt the standard technique (Shieber, 1986, p 29) of treating the list as a pair comprising a FIRST and a REST, where the FIRST is the first item of the list and the REST is a list comprising all but the first element. The second element is then the FIRST of the REST, the third element, the FIRST of the REST of the REST, and so on down the list until the REST value is a special symbol — we shall use NIL — which marks the end of the list. Thus ⟨a b c⟩ becomes the AVM in Fig. 9.4.

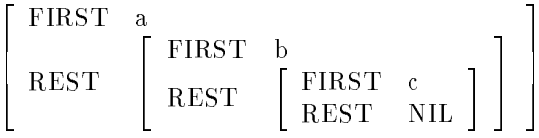


Figure 9.4: List structure

The rewritten, abbreviation-free AVM is shown in Fig. 9.5.

The DAG is drawn in Fig. 9.6. Arcs are labelled by attributes and some terminals, with values. The re-entrancies marked by the boxed variables in the matrix notation are illustrated directly, by arcs leading to the same node.

Finally, the thesis is not concerned with spoken language, and the formalisation in this chapter uses the spelt form of a word simply as an identifier, so WORD has been substituted for PHON.

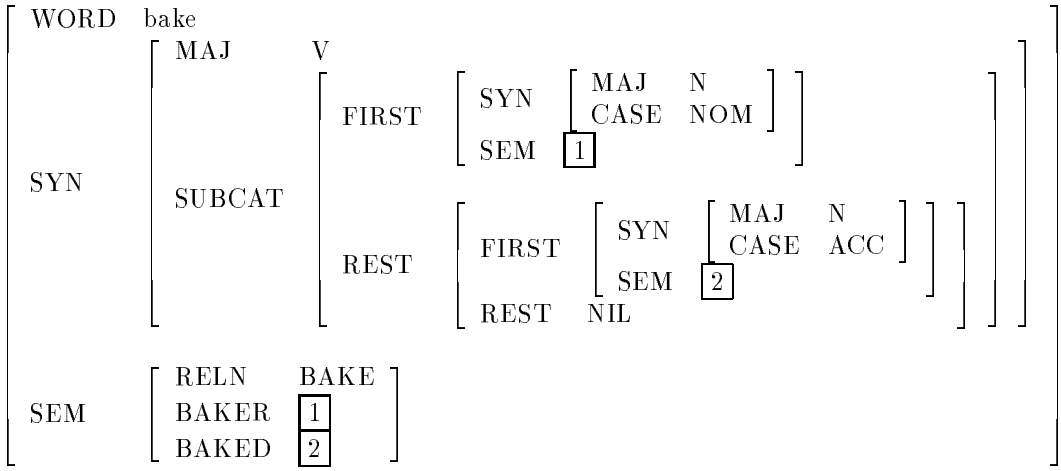


Figure 9.5: AVM for transitive *bake.*

The matrix is also equivalent to the following set of equations written in PATR:

Bake: <WORD> = bake
<SYN MAJ> = V
<SYN SUBCAT FIRST SYN MAJ> = N
<SYN SUBCAT FIRST SYN CASE> = NOM
<SYN SUBCAT FIRST SEM> = <SEM BAKER>
<SYN SUBCAT REST FIRST SYN MAJ> = N
<SYN SUBCAT REST FIRST SYN CASE> = ACC
<SYN SUBCAT REST FIRST SEM> = <SEM BAKED>
<SYN SUBCAT REST REST> = NIL
<SEM RELN> = BAKE

Where the AVM contains a pair of boxed variables, the two AVM paths involved are contracted into one PATR equation. The equals sign for these re-entrancies has a different semantics to the basic case where the item on the right hand side of the equation is a value. It entails that, for any attribute lists A, B and C and value V, if we have A = B and AC = V (where AC is the concatenation of A and C), then we have a theorem BC = V.

Figure 9.6: DAG for transitive *bake*.

9.3

optional arguments is for two predicates to have the same name but different arity, with the lower-arity predicate being defined as the higher-arity predicate with one of its arguments set, and this is the strategy we adopt. The argument is 'set' through existential quantification, so the alternation is as for the relation between transitive and unspecified-object *bake*. Thus we shall take the base form to be the "clear X of Y" one, represented as an AVM in Fig. 9.9.

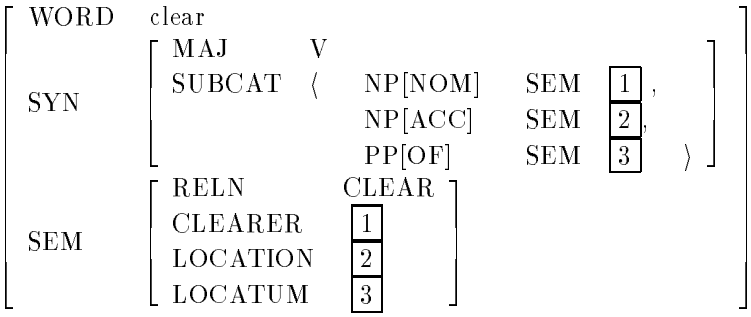


Figure 9.9: AVM for ditransitive 'of' *clear.*

A prepositional phrase abbreviation, PP[OF], has been introduced to stand for the AVM in Fig. 9.10. The PFORM feature is borrowed from GPSG (Gazdar, Klein, Pullum, & Sag, 1985) and

9.3.

9.3.4 *Melt* alternations as manipulations of AVMs

With *bake* and *clear*, we were able to represent the extended senses directly in terms of the same predicate that applied in the base sense. But, for *melt*, the intransitive ("The ice melted") is basic and the transitive ("Maria melted the ice") is extended, and it is not possible to define the extended sense directly in terms of the basic.

The transitive can be paraphrased using *cause*, "Maria caused the ice to melt"; the alternation is called the 'causative'. It is clearly closely related to the ergative, and it would be possible to treat the transitive form as basic, with the ergative alternation applying. That route has not been followed for two reasons. Firstly, *melt* is a member of a class of physical-process verbs, also including *evaporate, freeze, dissolve, sublime* and *coalesce*. They all clearly have intransitive senses. They all might, in the right setting, be used transitively, but in cases such as *coalesce* the transitive is not a standard use and it would patently be inappropriate for it to be treated as a base form. If we are to stand by the intuition that these verbs form a class, and all participate in the same alternation, then all must have an intransitive base form.

Secondly, transitive *melt* introduces an aspect of the path not

We have seen how the former may be addressed: let us move on to the latter.

We now need to translate AVMs into DATR. As far as possible, for each PATR equation we shall have a DATR equation which looks very similar. Let us return to *bake*. We translate it as follows.

```
Bake:<word> = bake
     <syn maj> = v
     <syn subcat fi syn maj> = n
     <syn subcat fi syn case> = nom
     <syn subcat re fi syn maj> = n
     <syn subcat re fi syn case> = acc
     <syn subcat re re> = nil
     <sem reln> = bake
     <syn subcat fi sem binding> = v1
     <sem baker binding> == v1
     <syn subcat re fi sem binding> == v2
     <sem baked binding> == v2.
```

DATR paths must be associated with nodes, so a node for the paths to be located at has been introduced. FIRST and REST have been shortened to `fi` and `re`. Upper case has been changed to lower case and PATR or AVM paths have become DATR paths.

DATR is not a unification formalism, and all the theory will do in relation to re-entrancies will be mark them with matched pairs of variables. Another module working on DATR output will be needed to interpret the matched pairs as re-entrancies. We introduce the feature `binding` for the variables to be the value of. If there were no such feature, so we had:

```
Bake: <syn subcat fi sem>    == v1
      <sem baker>            == v1.
```

then the inheritance of values for other paths starting `<syn subcat fi sem>` or `<sem baker>` from points above `Bake` in the inheritance hierarchy would be overridden. Also, as we shall see, the `binding` feature makes it possible to use the fact that a semantic argument has an existential-quantification binding to override the default that it is bound to a complement. The only kind of re-entrancy which occurs in the following fragment serves to unify complements with semantic arguments occurring along paths starting SEM. Thus all re-entrancy will be to the first second, third *etc.* member of the subcat list, or complement. The atoms denoting these positions will always be `v1, v2, v3` *etc.*, respectively.

$$\left[\begin{array}{ll} \text{WORD} & \text{melt} \\ \text{SYN} & \left[\begin{array}{lll} \text{MAJ} & \text{V} & \\ \text{SUBCAT} & \langle \ \text{NP[NOM]} \quad \text{SEM} \ \boxed{1}, & \\ & \quad\ \text{NP[ACC]} \quad \text{SEM} \ \boxed{2} & \rangle \end{array}\right] \\ \text{SEM} & \left[\begin{array}{ll} \text{RELN} & \text{CAUSE} \\ \text{CAUSER} & \boxed{1} \\ \text{CAUSED} & \left[\begin{array}{ll} \text{RELN} & \text{MELT/1} \\ \text{MELTED} & \boxed{2} \end{array}\right] \end{array}\right] \end{array}\right]$$

Figure 9.15: AVM for causative *melt*.

and:

Bake is a cooking verb, and cooking verbs are, in the base case, transitive change-of-state verbs. Thus `Bake` inherits, by default, from `COOKING-VB` which inherits from `C-OF-S` (for 'change of state') and then from `TRANSITIVE`, so acquiring the default specifications for semantic features for its subject and object, and the re-entrancies between subject and first argument, and object and second argument. The DATR fragment now represents all the information in the DATR lexical entry for *bake* presented above, and proto-agent and proto-patient specifications in addition.

Many change-of-state transitives, *bake* amongst them, can undergo the ergative alternation to become change-of-state intransitives, or 'physical process' verbs. Queries regarding the ergative forms will have paths beginning `alt erg`. The semantics of the ergative will be as for the base form. The third line of the `C-OF-S` node tells us, with the double-quotes, to inherit the ergative's semantics from the semantics of the global-context node, which will be the node for the base form of the verb. The two further specifications are that the first argument is existentially quantified, and the second unifies with the first complement via `UNI:<first>`.

In all other matters, the ergative form is diverted to inherit from a node for physical-process intransitives:

```
PHYS-PROC:<> ==  VERB
    <sem args fi semfeats> == PATIENT:<>.
```

The first semantic argument of a physical-process intransitive typically has proto-patient semantic features and otherwise inherits from `VERB`. We now have almost all the information needed to build the lexical entry for ergative *bake*. One item we do not yet have is the intuitively obvious fact that the `word` for the alternate form is the `word` for the original. This is true by definition for all alternate forms. All alternate forms will eventually have all their `alt` x prefixes stripped and inherit from `WORD-CLASS` at the top of the tree. So we add the following line:

```
WORD-CLASS:<word> == "<word>".
```

Now all alternate forms will inherit their `word` from the `word` at the global context node, which will always be the node for the base form.

Many cooking verbs undergo the 'unspecified object' alternation, for which we shall use the label `unspec`. All information relating to this form is gathered at an `UNSPEC` node:

```
UNSPEC:<>   ==  VERB
    <sem> == "<sem>"
    <sem args re fi binding>  == ex-q.
```

This simply states that the form is a standard intransitive, with the semantics of the base form except that the second argument is existentially quantified. Cooking verbs with `alt unspec` prefixes are diverted here with:

```
COOKING-VB:<alt unspec>  == UNSPEC:<>.
```

9.5.3 *Melt* and causatives

Melt is a physical-process verb which has a causative form. The ergative alternation led from `C-OF-S` to `PHYS-PROC`. This makes a similar journey in the opposite direction, from `PHYS-PROC` to `CAUSE` and then `TRANSITIVE`. The alternation label is `cause`.

```
Melt:<> == PHYS-PROC
    <sem pred> == melt /1
    <word> == melt.
```

```
PHYS-PROC:<> ==  VERB
```

```
Bake/erg:  <> == "Bake:<alt erg>".
```

This says not only that ergative *bake* inherits all its values from the base form of *bake* with `alt erg` prefixed, but also that the global context is changed so that the stored node is `Bake`, rather than `Bake/erg`, so the behaviour is exactly as if the query had been made at `Bake`. For every alternation, it is technically a trivial matter to produce such an alternation-specific node and it could be done automatically (though see section 8.7 for the pitfalls); in this chapter, such nodes will be freely created (with the *word/alternation-label* naming convention as here), and queried, without further discussion.

9.6.2 Which paths to query? A query grammar

A system using the lexicon will need some way of knowing

Figure 9.16: FSTN for interesting paths

gives the following output:

```
Bake: <lexical> = true.
Bake: <word> = bake.
Bake: <syn maj> = verb.
Bake: <syn subcat fi syn maj> = n.
Bake: <syn subcat fi syn case> = nom.
Bake: <syn subcat fi sem binding> = v1.
Bake: <syn subcat re fi syn maj> = n.
Bake: <syn subcat re fi syn case> = acc.
Bake: <syn subcat re fi sem binding> = v2.
Bake: <syn subcat re re> = nil.
Bake: <sem pred> = bake /2.
Bake: <sem args fi binding> = v1.
Bake: <sem args fi semfeats volition> = yes.
Bake: <sem args fi semfeats sentient> = yes.
Bake: <sem args re fi binding> = v2.
Bake: <sem args re fi semfeats changes-state> = yes.
Bake: <sem args re fi semfeats causally-affected> = yes.
Bake: <sem args re re> = nil.
```

9.6.3 Predicate names

Lexical entries can be kept smaller if we observe the redundancy associated with the `word` and `pred` paths. The predicate is built, in the default case, by taking the word and adding `/1` for an intransitive and `/2` for a transitive. The default applies to all forms seen so far except causative change-of-states. So we add the following lines:

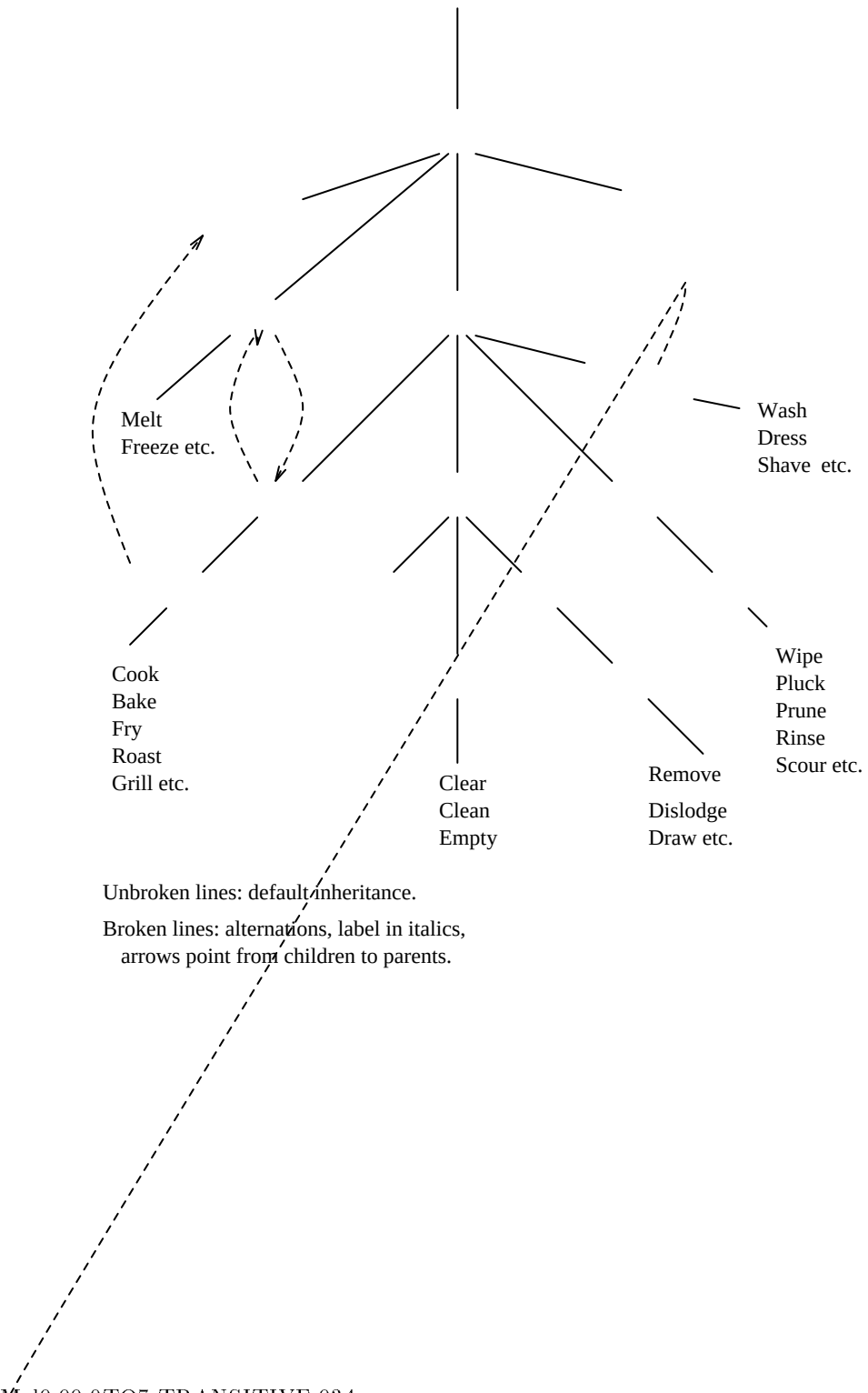
```
VERB:       <sem pred> == "<word>" /1.
TRANSITIVE: <sem pred> == "<word>" /2.
```

It is no longer necessary to state a `sem pred` equation at any lexical node.[8]

Predicates will be DATR sequences. This is not, of course, to say they should be treated as anything other than atoms by a parser or other program which uses the output of the lexicon, and it may be desirable to delete the white space between the elements of these DATR sequences prior to using them as DAG values, but that is an issue for the client system rather than the DATR lexicon.

It is to be noted that, while this is a saving on the amount of typing involved in inputting new lexical entries, it is not of theoretical interest. No generalisation relating words, which are linguistic entities, with predicates, which are semantic ones, has been captured. The generalisation is a trivial one relating words to *names* of predicates.

9.7 he LR verb class taxonomy


Melt
Freeze etc.
Wash
Dress
Shave etc.
Cook
Bake
Fry
Roast
Grill etc.
Clear
Clean
Empty
Remove
Dislodge
Draw etc.
Wipe
Pluck
Prune
Rinse
Scour etc.
Unbroken lines: default inheritance.
Broken lines: alternations, label in italics,
arrows point from children to parents.

9.7.

We need to swap the `of` for `from`. This is the `pform` for all 'remove' verbs, whether 'remove' is a base or extended form, so can be stated at `REMOVE`.

The DATR is:

```
SURFACE-CONTACT:<alt conative>    == CONATIVE.

CONATIVE:<>                       == TRANSITIVE
     <sem pred>                   == (conative /1)
     <sem args fi>                == "<sem>"
     <sem args re>                == NIL:<>
     <syn subcat re fi syn>       == PP:<>
     <syn subcat re fi syn pform> == at.
```

The last two lines express the syntactic difference. A conative form still has two complements so inherits from `TRANSITIVE`, but we override the default that complements are NPs by stating that the second complement is a PP with pform *at*.

Not all 'wipe' verbs participate readily in the conative alternation. The issue is as addressed in the previous chapter. We have set up a fragment to indicate which verbs *may* participate in which alternations, but not which *do*. There were two mechanisms presented in the last chapter: a negative one, of building an `altlist` stating which alternations may apply to a word (and by implication which may not), and a positive one of stating where a usage-type was attested. Both mechanisms are also applicable to the fragment of the verb lexicon.

The quotation from LR claims the conative alternation may be associated with any verbs with both 'contact' and 'motion' meaning components. So a conative sense is to be added to the `altlists` of just those verbs. Meaning components of verbs are distinct from the semantic features we have considered so far. The latter, such as 'volition' and 'causally-affected', apply or fail to apply to entities, prototypically physical objects such as people and chairs, which are standardly linguistically realised as nominals. In the verbal lexicon, they relate to the arguments of verbs. By contrast, 'contact' and 'motion' apply or fail to apply directly to verbs rather than to their arguments. So, in addition to `semfeats`, we introduce a new feature, `vsemfeats` for components of verbal meaning simpliciter. So:

```
Rub:<sem vsemfeats contact> = yes.
Rub:<sem vsemfeats motion> = yes.
Move:<sem vsemfeats contact> = no.
Move:<sem vsemfeats motion> = yes.
Touch:<sem vsemfeats contact> = yes.
Touch:<sem vsemfeats motion> = no.
```

`conative` should be added to the `altlist` for any transitive verb if and only if the answer to both these queries is `yes`. The mechanism for doing this in DATR, which involves more sophisticated DATR programming techniques than we have used so far, is presented in Appendix C.

It is not at all clear what we might include in the inventory of `vsemfeats`. The analysis has freely used an inventory of the semantic features that verbs expect of their arguments.[10] But where LR say they consider it their project to uncover "syntactically relevant components of verb meaning" (p 123) they are concerned with aspects of meaning which relate directly to the verb rather than to its arguments, and it is these 'verb meaning components' or 'verbal semantic features' that the attribute `vsemfeats` is provided for. LR mention several candidate `vsemfeats`, 'contact' and 'motion' among them, but their analysis falls far short of providing a motivated inventory of primitive features. This is of course scarcely surprising. A large number of critiques of Katz & Fodor (1963)'s decompositional semantics point to the theoretical and practical

[10] Dowty talks of entailments that a verb licenses about its arguments rather than semantic features: the distinction is not important here.

9.8.2 Syntax, semantics and circularity

LR's enterprise runs a risk of circularity which the authors do not mention. Their goal is to establish which 'components of meaning' have linguistic significance. The risk is that an alternation not only provides evidence of linguistic significance, but also proves criterial for whether a given 'component of meaning' is present. For example, *shut* is presented as belonging to the same class as *thicken* (p 134): they are both change-of-state verbs. But what is the evidence that they share components of meaning? The clearest evidence is syntactic: they are both related to adjectives and undergo a characteristic alternation. LR do not discuss how 'components of meaning' are to be identified independently of syntactic, and in particular alternation-based, criteria. So their claim to be unearthing correlations between syntactic and semantic domains is a shallow one: to be strengthened, the philosophically tortuous question of non-syntactic criteria for identifying

First, HPSG-style verbal lexical entries, and the mappings between them corresponding to alternations, were described. But at this stage, the generalisations were not captured. So then these entries were translated into DATR, and arranged into a taxonomy so an alternation only needed expressing once, at a non-terminal node from which the verbs to which it applied would inherit. The theory was developed to succinctly represent lexical entries for seven classes of base verbs and eight alternations applying to or between one or more of the classes. Information about syntax, semantics, and patterns of polysemy was concisely expressed in a manner both theoretically and computationally appealing.

Finally some constraints on the approach were discussed. The articles and the formal theory assumed a simple distinction between what was, and what was not, part of a language. An important area for the development of the approach is to establish ways in which statistical information, regarding more and less likely uses of words in a language, can be incorporated.

Chapter 10

Conclusion

What is polysemy? How is it that language-users can effortlessly comprehend and generate novel uses of words? How might natural language processing computer systems deal with multiple

The fifth who chanced to touch the ear
Said, "Even the blindest man
Can tell what this resembles most;
Deny the fact who can,
This marvel of an elephant
Is very like a fan."

The sixth no sooner had begun
About the beast to grope
Than, seizing on the swinging tail
That fell within his scope,
"I see," cried he, "the elephant
Is very like a rope."

— John Godfrey Saxe

Polysemy, like the elephant, may at first encounter seem like a variety of things: like homonymy, ellipsis, metaphor; like syntactic variation, collocation, pragmatic reasoning. The question, "What does it mean to say a word has many meanings?", may be addressed from any of these angles. This thesis has aimed to show the beast in its entirety.

The empirical studies looked directly at polysemy as found in a dictionary, and the relations between that and the ways words were used in a corpus. An uncritical consideration of a dictionary might give rise to the following thought. It lists words, thereby providing a key to a set of entities in the language. Likewise, it lists word senses, so must also be providing a key to another set of entities. The dictionary studies set this misconception to rest. Different words are different because they have different spellings and sounds. There is no comparable fact of the matter for determining what makes a word sense different.

The two studies exposed a range of phenomena involving words having a variety of uses, and a corresponding range of lexicographical devices. We identifieov1499i12Td[(Can)-000(corres5(to)]TJ98.41uerto):12T

theories are of course small fragments, but may serve as prototypes for larger-scale projects in formal lexicography.

10.4 Further work

The thesis suggests several avenues for further research.

10.4.1 Evaluation of disambiguation systems

senses. Thus if 80% of occurrences of both *melt* and *freeze* are intransitive and 20% causative, we might expect the same to hold for *evaporate* and other physical-process verbs where, owing to lower absolute frequencies, we have no direct evidence. The 80:20 ratio could then be stated at the higher node in the inheritance hierarchy. There is a host of difficult questions regarding how such relative frequencies might be calculated, where they might be inherited, and what they would be good for. In particular, when should a word be assumed to follow the pattern for its semantic field and when must it to be treated as a special case? If those questions can be answered, then scant information on relative frequencies of different senses could be aggregated and used to contribute to our understanding of the behaviour of whole classes of words. Such reasoning is likely to play a major role in our understanding of the lexicon in the future.

10.5 o conclude: summary of principal contributions

The thesis, then, has contributed to our understanding of polysemy in a number of ways. To finish, we restate four principal conclusions and contributions.
The thesis:

- shows the Bank Model to be fatally flawed;
- presents the SFIP criterion, describing when it is appropriate for a paper dictionary to list a usage-type, and describes the relation of polysemy to the four phenomena it falls between: homonymy, collocations, alternations and analogy;
- presents formal theor9utions.

Appendix A

Words examined in matching study

Showing: words examined; parts of speech; number of senses for each part of speech (excluding senses for collocations and including, for nouns, any senses specific to the plural); and in the last column, '1' if the word had only one sense for each part of speech, 'N' if every usage could be

Word & forms	POS	No. of senses	Result
credit/s	n, v	7, 2	Y
criminal/s	a, n	3, 1	Y
critical	a	3	Y
curious	a	2	Y
custom/s	n	6	N
dangerous	a	1	1
decide/s/ed/ing	v, a	3, 2	Y
delightful	a	1	1
derived	a, v	0[2], 3	Y
design/s	v, n	2, 6	Y
destroy/s/ed/ing	v	2	N
detail/s	n	2	Y
distinction/s	n	3	Y
divine	a, v	2, 2	N
don/s/ed/ing	n, v	1, 1	1
dust/s/ed/ing	n, v	5, 2	Y
eleven/s	n	2	N
embassy/s	n	1	1
emphasis/es	n	1	1
energy/s	n	3	Y
ensure/s/ed/ing	v	1	1
enter/s/ed/ing	v	7	Y
entrance/s	n	3	Y
escape/s/ed/ing	n, v	2, 3	Y
establishment/s	n	4	Y
evil/s	n, a	1, 2	N

Word & forms	POS	No. of senses	Result
exchange/s	n, v	4, 1	Y
exciting	a	1	1
exercise/s/ed/ing	n, v	5, 3	Y
expensive	a	1	1
explain/s/ed/ing	v	2	Y
factory/s	n	1	1
farming	n	1	1
fashion/s	n, v	3, 1	Y
favourite/s	a, n	1, 3	Y
federal	a	2	Y
feed/s/ed/ing	n, v	4, 5	Y
flights	n	7	Y
football/s	n	4	Y
formal	a	4	Y
frame/s	n, v	6, 3	Y
friendship/s	n	2	Y
gallery/s	n	4	Y
gas/s/ed/ing	n, v	7, 2	Y
generation/s	n	4	Y
gift/s	n	3	Y
guest/s	n	4	Y
herring/s	n	1	1
hit/s/ing	n, v	5, 5	Y
ideal/s	n, a	2, 3	Y
image/s	n	6	Y

[2] no dictionary entry for adjectival form

Appendix B

Matching study: examples of misfits

The following are examples of citations which cannot be satisfactorily classified as one and only one of the dictionary senses of the word. Words have been selected to demonstrate a variety of kinds of cases.

arrive *v* [I] **1** to reach a place at the end of a journey: *We arrived home safely.* | *What time does the plane arrive in New York?* —compare DEPART (1) **2** to come to a place, esp. by arrangement: *Shall we start now, or shall we wait for the others to arrive?* **3** to be brought or delivered to a place: *Has the post arrived yet?* | *I'm still waiting for those books I ordered to arrive.* **4** to happen as expected or arranged; come: *At last the great day arrived.* | *Her baby arrived* (=was born) *yesterday.* **5** to win success: *They felt they had really arrived when they made their first record.*

1	barge. When the American symphony orchestra	arrive in Marlow on July 15 they will use an 18-foot
2	climax in the island orgy. Here, the guests	arrive in ghost-like yachts, the wildly flapping
3	3,000 men, who must take about a fortnight to	arrive. If the UN forces were thick enough on the
4	are things wrong with the film, but the print	arrived from the cutting room only a few hours before
5	him of the Vienna outcome. Last night Mr Rusk	arrived in London in time to join the Buckingham

Comments:

colour *BrE* ‖ **color** *AmE n* **1** [U] the quality in objects which allows the eyes to see the difference between (for example) a red flower and a blue flower when both are the same size and shape: *The book has illustrations in colour.* | *These insects can change colour.* | *a colour television* **2** [C] red, blue, green, black, brown, yellow, white, etc.: *"What colour is this paint?" "It's red."* | *"What colour did you paint the door?" "I painted it red."* **3** [S;U] the general appearance of a person's skin, esp. as this shows the state of their health: *He lost colour* (=became pale) *during his illness.* | *The fever gave her a* **high colour**. (=a lot of colour) | *The cold wind brought colour to her cheeks.* (=made them red) **4** [C] the colour of a person's skin showing which race they belong to: *people of all colours* (=black, brown, white, etc.)—see also COLOURED **5** [U] details or behaviour of a place, thing or person, that interest the mind or eye and excite the imagination; character: *She loved the life, noise and colour of the market.* | *The lecturer told a few jokes and anecdotes to* **add colour to** *his talk.* —see also LOCAL COLOUR **6 give/lend colour to** to make (something, esp. something unusual) appear likely or true: *Her wet hair lent colour to her claim that she had fallen into the lake.* **7 off colour** *infml* not in good health: *You look a little off colour today.* **8 see the colour of someone's money** *infml* to have clear proof that someone has enough money to pay:

3 [U] fine powder made of small pieces of the stated substance: *gold dust* | *coal dust* **4** [U] *lit* the earthly remains of bodies once alive: *the dust of our ancestors* **5** [S] an act of dusting: *I gave the living room a quick dust.* **6 kick up/raise a dust (about)** *infml* to argue and shout (about) **7 when the dust has settled** *infml* when the confusion is over —see also DUSTY, **bite the dust**

1	in the pits. In South Wales 346 died from	dust in 1959 alone. Miners who had worked during
2	who had worked during the long period of '	dust-approved conditions,' including younger men
3	there will be resistance.' 346 in a year.	Dust kills many more people than gas, bad roofs,
4	this combination. The body returns to the	dust, the life returns to God, and the spirit disappears.
5	that his seed should be like the stars, the	dust and the sand that can not be numbered, we know
6	holes in the drill body must be kept free from	dust, screws should be checked for tightness regularly,
7	mild steel and other softish metals splinters and	dust are a sign that more pressure is required, so

Comments: Similar to 'chapel' above, except that in this case, some dust is equally dust/1, dust/2 and dust/4, which is unlike 'chapel' since (almost) every individual chapel is either a chapel/1, chapel/2 or chapel/3. Note that, firstly, although case 1 refers to coal dust, it would be inappropriate to classify it as sense 4 since that requires that the substance (i. e. coal) is stated, as in the example noun-noun compounds. Secondly, case 4 alludes strongly to s4 but the dust that the body returns to (in this citation) is not the dust that the body turns into; the sense 4 figure of speech is implied but not employed.

image *n* **1** [C (**of**)] a picture formed in the mind: *She had a clear image of how she would look in twenty years time.* **2** [C] a picture formed of an object in front of a mirror or LENS, such as the picture formed on the film inside a camera or one's REFLECTION in a mirror **3** [C]this initiase, someit

Appendix C

Altlists and the conative in DATR

As discussed in section 9.7.3, LR claim the conative alternation is associated with verbs with both 'contact' and 'motion' meaning components. So a conative sense is to be added to the `altlists` of just those verbs. Using the hierarchy built in Chapter 9, we add `Rub`, `Move` and `Touch`, the salient `vsemfeats` specifications, and `altlist` values capturing alternations already discussed, to the theory as follows:

```
Rub:<>                          == WIPE.

WIPE: <sem vsemfeats motion>    == yes
      <altlist>                 == remove put SURFACE-CONTACT.

SURFACE-CONTACT:
      <sem vsemfeats contact>   == yes.

Touch:<>                        == SURFACE-CONTACT
      <sem vsemfeats motion>    == no.

Move: <>                        == TRANSITIVE
      <sem vsemfeats contact>   == no
      <sem vsemfeats motion>    == yes
      <altlist>                 == refl TRANSITIVE.
```

(`Move` undergoes the reflexive alternation, "I moved" meaning "I moved myself".) From this we want to derive that `Rub` but not `Move` or `Touch` has `conative` on its `altlist`.

As in the DATR for gathering collocations in Chapter 8, we gather alternations in the `altlist` by adding members to a sequence as we inherit up the hierarchy. `Rub`, `Move` and `Touch` all inherit, directly or indirectly, from `TRANSITIVE`, so the rule for adding `conative` to the `altlist` is stated there. Any other alternations —unconditional or, like `conative`, conditional— which apply to all transitives will also be stated there, and after gathering alternations from `TRANSITIVE`, an `altlist` query will proceed to see whether there are any more to be gathered from still higher up the hierarchy, from `VERB`. `VERB` is to be found at the end of the `TRANSITIVE:<altlist>` line below (and several lines of the proof) because, after determining whether `conative` is to be added to the `altlist`, DATR will look to see whether there are any more additions to be made at `VERB`.

We add `conative` if and only if, when we go back from `TRANSITIVE` to the base node and ask the two queries, we get `yes` twice. The truth-table aspect of the problem is identified at the `AND` node with the path prefix `truth-table`.

When introducing the `altlist` in section 8.6 we noted that it was a second order feature, conveying a different kind of information to that in the remainder of the theory. Now, there is also a technical difference. In the main theory, a negative result causes a query to fail, but in the `altlist` part of the theory, a negative result is represented as the empty sequence.

References

Aarts, J. & Meijs, W. (1990). *Theory and Practice in Corpus Linguistics.* Rodopi, Amsterdam.

Ahlswede, T. E. & Evens, M. (1988). Parsing versus text processing in the analysis of dictionary definitions. In *ACL Proceedings, 26th Annual Meeting*, pp. 217–224 New York.

Ahlswede, T. E. (1985). A linguistic string grammar for adjective definitions in w7. In Williams, S. (Ed.), *Humans and Machines: 4th Delaware Symposium on Language Studies*, pp. 101–128. Ablex, Norwood.

Amsler, R. A. (1980). *The Structure of the Merriam-Webster Pocket Dictionary.* Ph.D. thesis, University of Texas at Austin.

Apresjan, J. D. (1974). Regular polysemy. *Linguistics, 142*, 5–32.

Atkins, B. T. S., Kegl, J., & Levin, B. (1986). Explicit and implicit information in dictionaries. In *Advances in Lexicography: Proc. Second Ann. Conf. of the UW Centre for the New OED*, pp. 45–65 Waterloo, Canada.

Atkins, B. T. S., Kegl, J., & Levin, B. (1988).

Boguraev, B. K. & Pustejovsky, J. (1990). Lexical ambiguity and the role of knowledge representation in lexicon design. In *COLING 90*, Vol. 2, pp. 36–41 Helsinki.

Boguraev, B. K. (1979). *Automatic Resolution of Word-Sense Ambiguity.* Ph.D. thesis, Cambridge University.

Bolinger, D. (1965). The atomization of meaning.

Cowie, A. P. & Mackin, R. (1975). *Oxford Dictionary of Current Idiomatic English. Volume 1: Verbs with Prepositions and Particles.* OUP, Oxford.

Cowie, A. P. et al. (1982). *Oxford Dictionary of Current Idiomatic English. Volume 2.* OUP, Oxford.

Cruse, D. A. (1986).

Francis, W. N. (1964). *A standard sample of present-day American English for use with digital computers. Report to the U.S. Office of Education on Cooperative Research, Project No. E-007.* Brown University, Providence.

Garside, R., Leech, G. N., & Sampson, G. (Eds.). (1987). *The Computational Analysis of English.* Longman, Harlow.

Garside, R. (1987). The claws word-tagging system. In Garside, R., Leech, G. N., & Sampson, G. (Eds.), *The Computational Analysis of English.* Longman, Harlow.

Gazdar, G., Klein, E., Pullum, G. K., & Sag, I. (1985). *Generalized Phrase Structure Grammar.* Blackwell, Oxford.

Gazdar, G. (1987). *The grammatical form of words: inflectional morphology and the lexicon.* ESRC RC/00/24/2003.

Geeraerts, D. (1990). The lexicographical treatment of prototypical polysemy. In Tsohatzidis, S. L. (Ed.), *Meanings and Prototypes: Studies in Linguistic Classification*, pp. 195–210. Routledge, London.

Gibbon, D. (1990). Prosodic association by template inheritance. In *Proc. Workshop on Inheritance in Natural Language Processing*, pp. 65–81 Tilburg. ITK.

Gibbs, R. (1983). Do people always process the literal meanings of indirect requests?. *Journal of Experimental Psychology: Learning, Memory and Cognition, 9*(3), 524–533.

Gibbs, R. (1984). Literal meaning and psychological theory. *Cognitive Science, 8*(3), 275–304.

Gibbs, R. (1986). Skating on thin ice: literal meaning and understanding idioms in conversation. *Discourse Processes, 9*, 17–30.

Givon, T. (1967a). Some noun-to-noun derivational affixes. Tech. rep. SP-2893, System Development Corporation, Santa Monica, California.

Givon, T. (1967b). Transformations of ellipsis, sense development and rules of lexical derivation. Tech. rep. SP-2896, System Development Corporation, Santa Monica, California.

Guthrie, L., Slator, B. M., Wilks, Y., & Bruce, R. (1990). Is there content in empty heads?. In *COLING 90*, Vol. 3, pp. 138–143 Helsinki.

Guthrie, J. A., Guthrie, L., Wilks, Y., & Aidinejad, H. (1991). Subject-dependent co-occurrence and word sense disambiguation. In *Proc. 29th Annual Meeting of the Association for Computational Linguistics*, pp. 00–00 Berkeley.

Halliday, M. A. K. (1966). Lexis as a linguistic level. In Bazell, C. E., Catford, J. C., & Halliday, M. A. K. (Eds.), *In Memory of J. R. Firth*, pp. 148–162. Longman, Harlow.

Hearst, M. A. (1991). Noun homograph disambiguation using local context in large text corpora. In *Using Corpora: Proc. Seventh Ann. Conf. of the UW Centre for the New OED*

Hirst, G. (1986). Why dictionaries should list case structures. In *Advances in Lexicography: Proc. Second Ann. Conf. of the UW Centre for the New OED* Waterloo, Canada.

Hirst, G. (1987). *Semantic Interpretation and the Resolution of Ambiguity.* CUP, Cambridge, England.

Hofland, K. & Johansson, S. (Eds.). (1982). *Word Frequencies in British and American English.* The Norwegian Computing Centre for the Humanities, Bergen, Norway.

Hopcroft, J. E. & Ullman, J. D. (1979). *Introduction to Automata Theory, Languages and Computation.* Addison-Wesley, London.

Ilson, R. & Mel'čuk, I. (1989). English *bake* revisited. *International Journal of Lexicography*, *2*(4), 325–346.

Jakobson, R. & Halle, M. (1956). *Fundamentals of Language.* Mouton.

Jenkins, E. A. (1990). Enhancements to the Sussex Prolog DATR implementation. In Evans, R. & Gazdar, G. (Eds.), *The* DATR *Papers*, CSRP 139. School of Cognitive and Computing Sciences, University of Sussex, Falmer, Sussex.

Jensen, K. & Binot, J.-L. (1987). Disambiguating prepositional phrase attachment by using on-line dictionary definitions. *Computational Linguistics*, *13*, 251–260.

Johnson-Laird, P. N. (1983). *Mental Models.* CUP, Cambridge, England.

Katz, J. J. & Fodor, J. A. (1963). The structure of a semantic theory. *Language*, *39*, 170–210.

Kilbury, J., Naerger, P., & Renz, I. (1991). DATR as a lexical component for PATR. In *European ACL, Fifth Conference* Berlin.

Klavans, J. L., Chodorow, M. S., & Wacholder, N. (1990). From dictionary to knowledge base via taxonomy. In *Electronic Text Research: Proc. Sixth Ann. Conf. of the UW Centre for the New OED*, pp. 110–132 Waterloo, Canada.

Klavans, J. (1988). COMPLEX: a computational lexicon for natural language systems. In *COLING 88* Budapest.

Krishnamurthy, R. (1987). The process of compilation. In Sinclair, J. M. (Ed.), *Looking Up: An Account of the COBUILD Project in Lexical Computing*, chap. 3. Collins, London.

Lakoff, G. & Johnson, M. (1980). *Metaphors We Live By.* University of Chicago Press.

Lakoff, G. (1987). *Women, Fire and Dangerous Things.* University of Chicago Press.

Leech, G. (1981). *Semantics.* CUP, Cambridge, England.

Lesk, M. E. (1986). Automatic sense disambiguation usingM.

Lyons, J. (1977). *Semantics.* CUP, Cambridge, England.

MacWhinney, B. (1989). Competition and lexical categorization. In Corrigan, R., Eckman, F., & Noonan, M. (Eds.), *Linguistic Categorization*, pp. 195–241. John Benjamins, Amsterdam.

Malakhovski, L. V. (1987). Homonyms in english dictionaries. In Burchfield, R. W. (Ed.), *Studies in Lexicography*, pp. 36–51. OUP, Oxford.

Pollard, C. & Sag, I. A. (1987). *An Information-Based Approach to Syntax and Semantics. Volume 1: Fundamentals.* CSLI Lecture Notes, No. 13. Chicago University Press, Chicago.

Pollard, C. & Sag, I. A. (1992). *An Information-Based Approach to Syntax and Semantics. Volume 2: Agreement, Binding and Control.* CSLI Lecture Notes. Chicago University Press, Chicago.

Pustejovsky, J. (1989). Current issues in computational lexical semantics. In *ACL Proceedings, 4th European Conference*, pp. xvii–xxiii Manchester.

Pustejovsky, J. (1991a). The generative lexicon. *Computational Linguistics, 17*

Stock, P. F. (1983). Polysemy. In *Proc. Exeter Lexicography Conference*, pp. 131–140.

Talmy, L. (1985). Lexicalisation patterns. *Language Typology and Syntactic Description*, *3*, 57–149.

Taraban, R., McDonald, J. L., & MacWhinney, B. (1989). Category learning in a connectionist model: learning to decline the german definite article. In Corrigan, R., Eckman, F., & Noonan, M. (Eds.), *Linguistic Categorization*, pp. 163–193. John Benjamins, Amsterdam.

ten Hacken, P. (1990). Reading distinction in machine translation. In *COLING 90* Helsinki.

Turner, L., Grover, C., & Briscoe, E. J. (1989). The syntactic regularity of english noun phrases. In *ACL Proceedings, 4th European Conference*, pp. 256–263 Manchester.

Vendler, Z. (1967). *Linguistics in Philosophy*. Cornell University Press, Ithaca, New York.

Veronis, J. & Ide, N. M. (1990). Word sense disambiguation with very large neural networks extracted from machine readable dictionaries. In *COLING 90*, Vol. 2, pp. 389–394 Helsinki.

Vossen, P., Meijs, W., & den Broeder, M. (1989). Meaning and structure in dictionary definitions. In Boguraev, B. K. & Briscoe, E. J. (Eds.), *Computational Lexicography for Natural Language Processing*, pp. 171–192. Longman, Harlow.

Walker, D. E. & Amsler, R. A. (1986). The use of machine-readable dictionaries in sublanguage analysis. In Grishman, R. & Kittredge, R. (Eds.), *Analysing Language in Restricted Domains*. Lawrence Erlbaum, Hillsdale, New Jersey.

Way, E. C. (1991). *Knowledge Representation and Metaphor*. Kluwer, Dordrecht.

Weinreich, U. (1980). *On Semantics*. University of Pennsylvania Press.

Whittemore, G., Ferrara, K., & Brunner, H. (1990). Empirical study of predictive powers of simple attachment schemes for post-modifier prepositional phrases. In *ACL Proceedings, 28th Annual Meeting*, pp. 23–30 Pittsburgh.

Wierzbicka, A. (1987). Boys will be boys: 'radical semantics' vs. 'radical pragmatics'. *Language*, *63*(1), 95–114.

Wiggins, D. (1971). On sentence sense, word-sense and difference of word-sense. towards a philosophical theory of dictionaries. In Steinberg, D. D. & Jakobitz, L. A. (Eds.), *Semantics: An Interdisciplinary Reader*. CUP, Cambridge, England.

Wilensky, R., Mayfield, J., Chin, D., Luria, M., Martin, J., & Wu, D. (1988). The Berkeley UNIX

Wilks, Y. A. (1975). A preferential pattern-seeking semantics for natural language inference. *Artificial Intelligence*, *6*, 53–74.

Wittgenstein, L. (1953). *Philosophical Investigations.* Macmillan, London.

Zernik, U. & Dyer, M. G. (1987). The self-extending phrasal lexicon. *Computational Linguistics*, *13*, 308–327.

Zgusta, L. (1971). *Manual of Lexicography.* Mouton, The Hague.