A CONSTRUCTIONAL APPROACH TO IDIOMS AND WORD FORMATION

A DISSERTATION
SUBMITTED TO THE DEPARTMENT OF LINGUISTICS
AND THE COMMITTEE ON GRADUATE STUDIES
OF STANFORD UNIVERSITY
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Susanne Z. Riehemann
August 2001

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

Thomas Wasow
(Principal Adviser)

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

Ann A. Copestake

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

Eve V. Clark

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

Arnold M. Zwicky

Approved for the University Committee on Graduate Studies:

Abstract

This dissertation explores a constructional approach to various aspects of grammar, in particular idioms and derivational morphology, within the Head-Driven Phrase Structure Grammar (HPSG) framework. The approach views these as complex patterns with sub-components, as opposed to separate pieces and ways for assembling them. For idioms this means focusing on the phrasal level and viewing idiomatic words as parts of idiomatic phrases; for morphology, it means looking at the word level and viewing stems and affixes as parts of complex words. Neither the idiomatic words nor the affixes have an existence outside of these complex patterns, and their meaning is instead associated with the larger unit. This is motivated by the fact that affixes cannot occur by themselves, and many idiom parts cannot occur with their idiomatic meaning outside of these patterns. Further motivations for the constructional approach are discussed, and it is shown that these motivations are analogous for idioms and derivational morphology. The dissertation also explores phenomena that straddle the lexical/phrasal boundary, such as compounds and separable prefix verbs in German, and shows that they can be handled constructionally as well.

The data in this dissertation were obtained through an extensive study of large corpora. This study shows that idioms frequently occur in non-canonical forms. For example, idioms involving noun phrases are sometimes used with a different specifier or an added adjective. Many idioms can also occur in passive or raising constructions, across relative clause boundaries, and in other variations. Such variations account for about 25% of the corpus examples of semantically decomposable idioms. The approach outlined in this dissertation handles these variations by specifying the relationship between parts of the idiom on a semantic as opposed to a syntactic level.

Acknowledgments

There are many people who helped make it possible for me to write this dissertation. First of all I would like to express my gratitude to my four committee members. I would like to thank my advisor, Tom Wasow, for always being available whenever I needed him even when he was on sabbatical, for his encouragement and his interest in my work, and for a commencement speech that touched my mother's heart. This dissertation would not exist without Ann Copestake, who first suggested expanding my idioms paper into a dissertation. The analysis in Chapter 5 is based on her work, and her comments lead to numerous improvements throughout the dissertation. It was a pleasure to have Eve Clark on my committee. She read all the drafts quickly, provided detailed feedback, and showed much interest in my work. I am grateful to Arnold Zwicky for showing such enthusiasm for my data, and for giving me self-confidence when I needed it.

I would also like to thank Ivan Sag, who provided detailed comments on an earlier draft of this dissertation, and from whom I have learned a great deal. His work on constructions in HPSG greatly influenced the approach in this dissertation. Special thanks go to Emily Bender, with whom I discussed many of the issues that arose in the course of writing my dissertation. She always took time for me, understood my questions immediately and was very helpful and supportive. It would have been so much harder without her!

I have also benefited from useful discussions with Farrell Ackerman, David Beaver, Dan Flickinger, Andreas Kathol, Rob Malouf, Chris Manning, Gert Webelhuth, the Berkeley-Stanford construction discussion group, and the audiences of my presentations at the University of Canterbury in New Zealand, at the 1997 HPSG conference

in Ithaca, WCCFL 18 in Tucson, and at the 2000 HPSG conference in Berkeley.

I am grateful to the Stanford Linguistics Department and the LinGO project at CSLI for the financial and other support they gave me during my studies. The research in Chapter 6 is based on work with Emily Bender (Riehemann and Bender 1999) which was supported by the NSF under grant number IRI-9612682. The research in Chapter 7 is based on Riehemann (1998), which in turn is based on my Tübingen University Master's Thesis supervised by Erhard Hinrichs and Marga Reis (Riehemann 1993).

I remain absolutely responsible for any errors in fact or analysis. Being listed above, including being on my committee, does not imply agreeing with everything I say in this dissertation.

Finally I would like to thank everyone who provided welcome distractions and emotional support during the dissertation writing stage: my parents Franz and Elisabeth Riehemann and parents-in-law Ely and Esther Zalta, Colin Allen, Sascha Brawer, Kathryn Campbell-Kibler, Itamar Francez, Ela Harrison Widdows, Anne MacDougall, Uri Nodelman, Emma Pease, Kyle Wohlmut, and my Hebrew teachers Orna Morad, Hannah Berman, and Naomi Burns. My husband Edward Zalta never stopped believing in me and was patient when my health was bad and the going was slow, understanding when I worked on a night schedule for months, and always concerned for my well-being. His companionship and support were invaluable.

Contents

Abstract **v**

Acknowledgments **vi**

1 Introduction **1**

1.1 Outline of the Dissertation 7

1.2 HPSG Background 7

1.2.1 Typed Feature Structures 8

1.2.2 Type Hierarchies 9

1.2.3 The Lexicon 10

1.2.4 MRS 11

1.2.5 An example 12

1.3 Preview of the Constructional Approach 25

2 Data that Approaches to Idioms Need to Capture **26**

2.1 Introduction 26

2.2 Need for Phrasal Pattern 27

2.2.1 Absence of Literal Meaning 27

2.2.2 Idiomatic Words Are Not Free 27

2.2.3 No Literal Parse 30

2.2.4 Restricted Flexibility of Some Idioms 31

2.2.5 Canonical Forms 32

2.2.6 Idiom Families 35

2.2.7 Locus for the Metaphorical Mapping 37

2.2.8 Locus for Semantics of Non-Decomposable Idioms 38
2.2.9 More than Head-Argument Relationships 38
2.2.10 Interaction of Idioms and Syntactic Constructions 47
2.2.11 Collocations 47
2.2.12 Psycholinguistic Evidence 49
2.3 The Problem of Variation 52
2.3.1 Variants Differing in Inflection 52
2.3.2 Open Slots 53
2.3.3 Modification 54
2.3.4 Passive 59
2.3.5 Topicalization 60
2.3.6 Distribution Over Several Clauses 60
2.3.7 Other Variations 61
2.3.8 Pronominal Reference 63
2.3.9 Incomplete Idioms 64
2.4 Summary 65

3 A Corpus Study of Canonical Forms and Variation 67
3.1 Methodology 68
3.2 Idioms Discussed in the Literature 73
3.2.1 Decomposable Idioms 73
3.2.2 Non-Decomposable Idioms 81
3.2.3 Summary 83
3.3 Idioms Relevant for Argumentation 83
3.4 Idioms Containing Non-Independent Words 91
3.5 A Random Sample of V+NP Idioms 95
3.5.1 Decomposable Idioms 99
3.5.2 Non-Decomposable Idioms 124
3.5.3 Summary 130
3.6 Collocations 130
3.7 Constructions 136

3.8 Comparison with Non-Idioms 138
3.9 Summary 148

4 Alternative Approaches to Idioms 150
4.1 Word-Level Approaches 151
4.1.1 Multi-Word Lexeme 151
4.1.2 Subcategorizing for the Phonology 153
4.1.3 Subcategorizing for the Syntax 155
4.1.4 LFG 159
4.1.5 Subcategorizing for the Semantics 161
4.1.6 GPSG - Partial Functions 165
4.1.7 Summary of Problems with Word-Level Approaches 166
4.2 Phrasal Approaches 168
4.2.1 (Partially) Fixed Phonology 168
4.2.2 (Partially) Fixed Syntax 169
4.2.3 TAG 171
4.2.4 (Partially) Fixed Semantics 173
4.2.5 Jackendoff 173
4.3 Pulman - Quasi-Inference 178
4.4 Summary 180

5 The Constructional Approach 183
5.1 Introduction 183
5.2 The Constructional Approach 184
5.2.1 The Status of Idiomatic Words 187
5.3 How the Approach Deals With the Problems 195
5.3.1 Variability 195
5.3.2 Properties Shared between Literal and Idiomatic Words 208
5.3.3 No Literal Interpretation 209
5.3.4 Restricting the Flexibility of Idioms 209
5.3.5 Idiom Families 210
5.3.6 Locus for the Metaphorical Mapping 210

5.3.7 Locus for Semantics of Non-Decomposable Idioms 211
5.3.8 More than Head-Argument Relationships 214
5.3.9 Canonical Forms 218
5.3.10 Syntactic Constructions 223
5.4 Alternative Variants of the Approach 225
5.4.1 Variant 1 225
5.4.2 Variant 2 226
5.4.3 Variant 3 227
5.5 Summary 228

6 The Interaction of Idioms and Constructions 230
6.1 Introduction 230
6.2 Interaction Data 230
6.3 Absolute Constructions 232
6.4 Predicative Idioms 235
6.5 A Constructional Analysis 236
6.6 Individual Systems 241
6.7 Summary 243

7 Derivational Morphology 244
7.1 Introduction 244
7.2 The Morphological Data 248
7.2.1 German bar-Adjectives 248
7.2.2 English able-Adjectives 252
7.2.3 German ig-Adjectives 254
7.2.4 Semi-Affixes 255
7.3 Previous Approaches 256
7.3.1 Lexical Rules 256
7.3.2 Word Syntax 256
7.4 Problems of these Approaches 259
7.4.1 Affixes and Stems are Not Free 259
7.4.2 Sub-Patterns Needed to Structure Lexicon 259

7.4.3 Additional Mechanism Redundant . . . 259
7.4.4 Sub-Patterns Needed for Subregular Productivity . . . 260
7.5 Outline of the Proposed Approach . . . 261
7.6 The Formal Approach . . . 263
7.6.1 A Hierarchy of bar-Adjectives . . . 263
7.6.2 The Productive Schema . . . 267
7.6.3 The Higher Level Generalizations . . . 269
7.6.4 Zero-Derivation . . . 272
7.6.5 Properties of the Approach . . . 273
7.7 Hebrew Derivational Morphology . . . 275
7.7.1 Non-Concatenative Morphology . . . 275
7.7.2 Sub-Productive Patterns . . . 276
7.8 Summary . . . 279

8 Analogy between Phrasal and Lexical Patterns 281
8.1 Shared Properties of the Approaches to Idioms and Morphology . . . 281
8.2 Between Lexical and Phrasal . . . 282
8.2.1 Compounding . . . 283
8.2.2 Separable Prefix Verbs in German . . . 288
8.3 Summary . . . 294

9 Conclusions 295
9.1 Further Work - Experience-Based HPSG . . . 297

A List of the Corpora Used 302

B List of the Dictionaries Used 304

List of Tables

2.1 Percentage of Idioms Modified by Adjectives 34
2.2 Stranded and Non-Stranded Occurrences in the NYT Corpus 46

3.1 Results from the Study of Idioms from the Literature 84
3.2 Results from the Study of Random V+NP Idioms 131
3.3 Results from the Study of Non-Idioms 148

4.1 Idiom Approaches and their Problems 181

6.1 Contrast Patterns Based on 14 Speakers. 232
6.2 Contrast Pattern for Speaker 1 242
6.3 Contrast Pattern for Speaker 2 242

7.1 Sub-Productive Derivational Patterns in Hebrew 277
7.2 Instrument Nominalization Comprehension Experiment 277
7.3 Instrument Nominalization Production Experiment 278

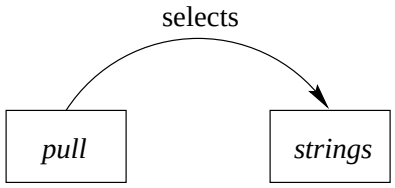
Chapter 1

Introduction

This dissertation studies various aspects of grammar, in particular idioms and derivational morphology, but also collocations, constructions, and compounding, and concludes that a constructional approach is needed for each of them. By 'constructional approach' I mean an approach that views things like idioms and derived words as complex patterns with sub-components, as opposed to separate pieces and ways for assembling them.

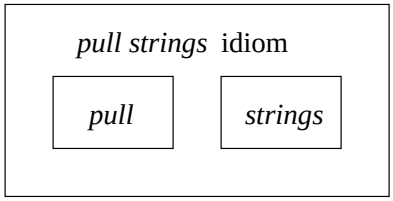
In general, idioms like *spill the beans* can be analyzed as special words with specific selectional restrictions (word-level approach) or as special phrases (phrasal or constructional approach). In word-level approaches, schematized in (1), there is a special lexical entry for *pull* that selects for the (special) word *strings* using the same mechanism that is usually used for subcategorization purposes, i.e. for stating how many complements a verb takes, whether they are NPs or PPs, etc. The meaning of the idiom is associated with the selecting word *pull*. Instances of this kind of approach differ in what exactly is specified about the idiomatic complements—its phonology, its syntactic identity, or its meaning.

(1)



In phrasal or constructional approaches, schematized in (2), the *pull strings* idiom

is a big phrase that contains within it the words *pull* and *strings*. In this dissertation I call such phrasal representations 'lexical entries', which I take to mean 'representation of a word or phrase specified in the lexicon'. Instances of this kind of approach differ in how they specify that the phrase contains these words, and how the words are related—via the phrase's phonology, via its syntactic daughters, or semantically.

(2) 

Various instances of these types of approaches are discussed in detail in Chapter 4.

I use the term 'idiom' to refer to an expression made up out of two or more words, at least one of which does not have any of the meanings it can have outside of the expression. As will become clear from the discussion below, this is not intended as an exact definition. A typical example of an idiom is *spill the beans*, where *beans* never means anything like 'secrets' in the absence of the word *spill*. These words are associated with the figurative meaning only as part of the whole idiom, and I call words which have this property 'idiomatic words'. Note that not all words that are part of an idiom have to be 'idiomatic words' in this sense. For example *miss* in *miss the boat* has the same meaning it can have in other expressions, so it is not an idiomatic word in this sense. But *miss the boat* is still an idiom according to my definition, because *boat* does not have the same meaning in other expressions. This definition of 'idiomatic word' is not as clear-cut as it may seem, because it may vary from speaker to speaker whether they think a word has the same meaning in other expressions. For example, some speakers use *spill* in expressions like *spill the secrets*, so it is not an 'idiomatic word' for them. By extension, the definition of idiom is not as clear as it may seem. For example, for some speakers words like *mold* in *break the mold* can have the same meaning in certain other expressions, such as *shatter the mold*, but not in general. For these speakers *break the mold* would not be an idiom according to the above definition. This does not match most people's intuition of what an idiom is, so the definition is not perfect. In fact Nunberg et al. (1994) argue

that no precise definition of idiom is possible, because 'idiom' is a fuzzy category that is as much defined by what is not an idiom as by what is. Prototypical idioms like *kick the bucket* have many properties, such as non-compositionality, (relative) inflexibility, figurativeness, etc., but none of these properties are present in all idioms. For more discussion of this see Chapter 3.

In this dissertation, the term 'collocation' is used for fixed expressions made up out of two or more words which do have one of the meanings they can have independently, and which combine compositionally, but which are conventionalized, i.e. established, in this particular combination, e.g. *bear the brunt of* or *hold one's turf*.[1] The term 'construction' is used for all syntactic constructions, from specific ones with non-compositional meanings like *What are your feet doing on the table?* (Kay and Fillmore 1999) to general syntactic constructions like the head-complement construction (see also Goldberg 1995, Zwicky 1994, and Sag 1997).

Furthermore, a distinction is made between semantically decomposable and non-decomposable idioms, following Nunberg et al. (1994).[2] Note that 'decomposable' has nothing to do with whether or not it is possible to guess the meaning of an idiom or its metaphorical motivation. Instead 'decomposable' is only intended to mean that parts of the meaning of the idiom are associated with parts of the idiom. Examples of decomposable idioms are *pull strings* and *spill the beans*, where *spill* roughly means 'reveal' and *beans* roughly means 'secrets'; while typical non-decomposable idioms are *kick the bucket* and *shoot the breeze*. *Kick the bucket* roughly means 'die', i.e. it is a one-place relation in which *the bucket* plays no role, and instead the meaning 'die' is associated with the whole idiom. Note that for the purposes of this dissertation it is not important whether all idioms can be classified into these two types without problems, although as discussed in Chapter 3 this poses a bit of a problem for the corpus study. As Wasow et al. (1983) also note, there might be some variation depending

[1] These are sometimes called 'idioms of encoding' (Makkai 1972:57).

[2] The terminology employed in that paper is different—they call decomposable idioms 'idiomatically combining expressions' and non-decomposable idioms 'idiomatic phrases'. I use this different terminology to emphasize the top-down perspective of being able to distribute the meaning over the parts of the idiom rather than the bottom-up one of combining the parts to build up the meaning. It also avoids potential confusion due to the fact that decomposable idioms are both idiomatic and phrases, although not completely fixed phrases.

on which paraphrase speakers think of, and perhaps not all speakers perceive these idioms the same way. But as long as idioms of both types exist, it is necessary to have an account of both types.

Almost all idioms are variable to some extent and cannot be seen as a simple fixed string of words. As Chapter 2 and Chapter 3 show in detail, the noun phrases in most idioms can vary in number and definiteness or include a different specifier, and the nouns can be modified by adjectives. For example, variations of the idiom *bury the hatchet* include *bury* ***their*** *hatchet****s***, and *bury the* ***legal*** *hatchet*. Syntactic variations like passivization, raising and distribution of the idiom parts over a main clause and subordinate relative clause are also possible, as for example in *The hatchet now* ***appears to have been buried*** *for good* and *Few folks ... speculated on* ***the hell*** *that would have been* ***raised*** *by George Steinbrenner*. Nevertheless, it is desirable to list an idiom only once, and use independently existing mechanisms of the grammar to derive the variations. In particular, inflectional information for idiomatic uses of words is identical to that of nonidiomatic uses, and should not have to be repeated. Therefore any approach has to establish some sort of link to the non-idiomatic lexical entries.

One reason why there is such a wide range of approaches in the literature is that they differ in their assumptions about what the data are. My corpus study in Chapter 3 shows that there is more variability than many linguists think, and than most phrasal approaches can handle. But there is less variability than word-level theories predict. In traditional approaches to idioms in the literature (e.g. Katz 1973, Chomsky 1980), their invariability was emphasized, and idioms were often inserted as fixed verbs in order to be compatible with assumptions about lexical insertion. Nunberg et al. (1994) showed that the parts of many idioms have meaning, and that this correlated with the syntactic variability of idioms. Some linguists have thought that handling such data requires a word-level approach. With the formal tools available at the time this may have seemed to be the only option. But there is a problem with such approaches in that they require a special mechanism to prevent *beans* with its idiomatic meaning from occurring elsewhere. This mechanism has to be general enough to deal with idioms with many words like *let the cat out of the bag*.

To my knowledge no such mechanism has been precisely described in the literature. Nunberg et al. (1994) show that parts of idioms are associated with parts of the idiomatic meaning, and that the relationship between idiom parts is semantic in nature. But they do not show there is a need for a word-level approach. As we will see in Chapter 5, a 'flat' semantics formalism that allows for underspecification, like MRS (see Section 1.2.4), makes it possible to represent idioms at the phrasal level, while still handling variability because argument taking and scoping are not represented as embedding.

The approach presented in this dissertation views both idioms and derived words as complex patterns with sub-components, as opposed to separate pieces and ways for assembling them. For idioms, this means focusing on the phrasal level and viewing idiomatic words as parts of idiomatic phrases. For morphology it means looking at the word level and viewing stems and affixes as parts of complex words. In this approach neither the affixes and stems nor the idiomatic words have an existence outside of these complex patterns.[3]

In the case of idioms, the main motivation is that words cannot occur in their idiomatic meanings outside of the idiom—some, like *dint* cannot occur outside of the idiom at all—and that idioms can involve more than just head-argument combinations of the kind usually specified via subcategorization. This is similar to the characterization of idioms as 'conventionalized complex expressions' in Everaert et al. (1995):

> In one sense they [idioms] are complex in that they consist of more than one word; in some other sense they are units. From a broader perspective, all authors in this volume agree that these complex units are syntactic expressions that exhibit lexical co-occurrence restrictions that cannot be explained in terms of regular rule-governed syntactic or semantic restrictions. (Everaert et al. 1995:3)

As is shown in detail in Chapters 2 and 3, there is also independent motivation

[3]One may think that words like *the* also do not exist outside of noun phrases. This could perhaps also be handled constructionally, although complex specifiers like *all the* complicated the matter somewhat. But the main difference is that the distribution of *the* is for the most part the same as that of other words of the category *determiner*, and syntactic theory has already developed mechanisms to ensure that words of this category only appear in the right contexts.

for the existence of the pattern, since speakers have knowledge of the canonical forms of idioms. Almost all idioms have a strongly preferred canonical form, which is not the case for non-idioms. For example speakers know that *the beans* in *spill the beans* is plural and definite and not modified by an adjective, while this is not the case for *reveal a/the secret(s)*. But the more abstract pattern is also needed to account for the productive variations of the idiom like passivized occurrences and occurrences where the idiomatic noun is modified by an adjective. As we will see in Chapter 5, this flexibility is achieved by specifying the relationship between the parts of the idiom on a semantic level. Also, in some semantically non-decomposable idioms such as *kick the bucket*, the phrasal pattern is needed to carry the semantics, because the meaning of the idiom cannot be distributed among its parts.

In the case of derivational morphology, the main motivation for the constructional approach is that the complex patterns are needed independently to structure the large inventory of lexicalized words. A separate lexical rule mechanism would be redundant because the same patterns can be used for productive word formation. The patterns are also needed to explain subregular productivity, and to account for non-concatenative morphology. Further motivation comes from the fact that affixes and stems are items which can never occur by themselves. All these motivations are parallel for idioms and morphology, although the emphasis is a bit different. Other similarities are discussed in Chapter 8. That chapter also explores further applications of this general approach to phenomena that straddle the lexical/phrasal boundary, such as compounds and separable prefix verbs in German. Jackendoff (1997) also sees this relationship between morphology and idioms, and has a chapter containing a similar approach for dealing with idioms. This is discussed in Chapter 4.

Another related framework is Construction Grammar (Fillmore and Kay 1997, Kay and Fillmore 1999, Fillmore et al. 1988, Goldberg 1995). However, it is different in several ways. My approach is more general in that it includes a 'constructional' approach to morphology as well as syntax. It also presents some of the issues, such as relating patterns, relating idiomatic and non-idiomatic words, and integrating phrasal and lexical meaning, in a more formally precise way, and accounts for the fact that words cannot appear with their idiomatic meaning outside the idiom. The approach

is more consistently 'constructional' in that it can use phrasal patterns even when variability needs to be dealt with.

1.1 Outline of the Dissertation

Chapter 2 discusses the data which approaches to idioms need to be able to handle, and which motivate the constructional approach. Chapter 3 contains the results of a corpus study which provides more information on canonical forms and the variability of idioms. Chapter 4 considers the range of possible approaches to idioms within the HPSG framework, and also discusses approaches from the literature. Chapter 5 presents the constructional approach, and explains how it can handle the data and solve the problems discussed in Chapter 2. Chapter 6 contains an application of the approach to the interaction of idioms and constructions, specifically predicative idioms and the *with* and *with*-less absolute constructions. This chapter is based on work with Emily Bender (Riehemann and Bender 1999). Chapter 7 presents a constructional approach to derivational morphology. It is based on Riehemann (1998), the main addition being a section applying the approach to Hebrew. Chapter 8 shows what the approaches to idioms and morphology have in common, and also examines two phenomena which are neither strictly 'lexical' nor 'phrasal': compounding and separable prefix verbs in German. The remainder of this introduction provides some background information about HPSG.

1.2 HPSG Background

The approach presented in this dissertation uses the HPSG framework (Pollard and Sag 1994, Sag and Wasow 1999). In this section I give a brief overview of the version of HPSG I assume, which is essentially the 'constructional' version of HPSG developed in Sag (1997) coupled with MRS (Copestake et al. 1999) for the semantic representations. It is also very similar to the version of HPSG implemented in the LinGO project.

HPSG is a linguistic theory that emphasizes the need for integration of partial

information from various linguistic (and in principle also nonlinguistic) sources in an order-independent way: the grammar is formulated as a declarative system of constraints. This makes it possible to use the same grammar for parsing and generation, i.e. language understanding and production. It is also important for my approach to idioms, because as we will see in Chapter 5 the idiomatic constructions are represented as constraints on phrasal types which contain pieces that cannot necessarily be integrated into the representation for the whole sentence at a particular point during processing. In the next few subsections I describe various aspects of the the HPSG framework in general terms, and then illustrate them using a concrete example.

1.2.1 Typed Feature Structures

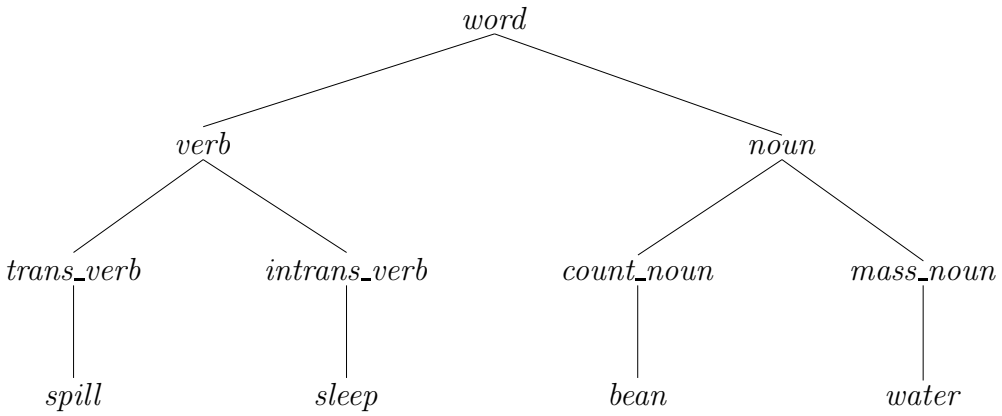
The fundamental objects of linguistic analysis in HPSG are signs, which allow for the parallel representation of phonological, syntactic, semantic, and other information. These signs are modelled by feature structures. Feature structures are described by AVMs (attribute-value-matrices), which are essentially sets of pairs of attributes and their values, although identity between the values of different features can also be expressed. The values can themselves be complex. The feature structures employed in HPSG are typed, where the type indicates what kind of object is being described. Features are defined as appropriate only for objects of certain types. Syntactic tree structure is encoded in the AVMs by means of daughter attributes such as HEAD-DAUGHTER (HEAD-DTR).

Two paths in a feature structure can lead to the same node, which is called structure sharing. It is indicated by boxed numerals in AVMs and involves token identity. If constraints are given in more than one place connected by such structure sharing, this information has to unify, i.e. it has to be consistent. Unification is a general method used for merging information from different sources. For example, the information stated in a lexical entry and that of a slot in a phrasal constraint in which it is to be inserted is combined by unification. If there is a conflict, unification fails. This ensures, for example, that it is not possible to insert a *verb* as the HEAD-DTR of a *noun_phrase* construction, which specifies that its HEAD-DTR must be a *noun*.

1.2.2 Type Hierarchies

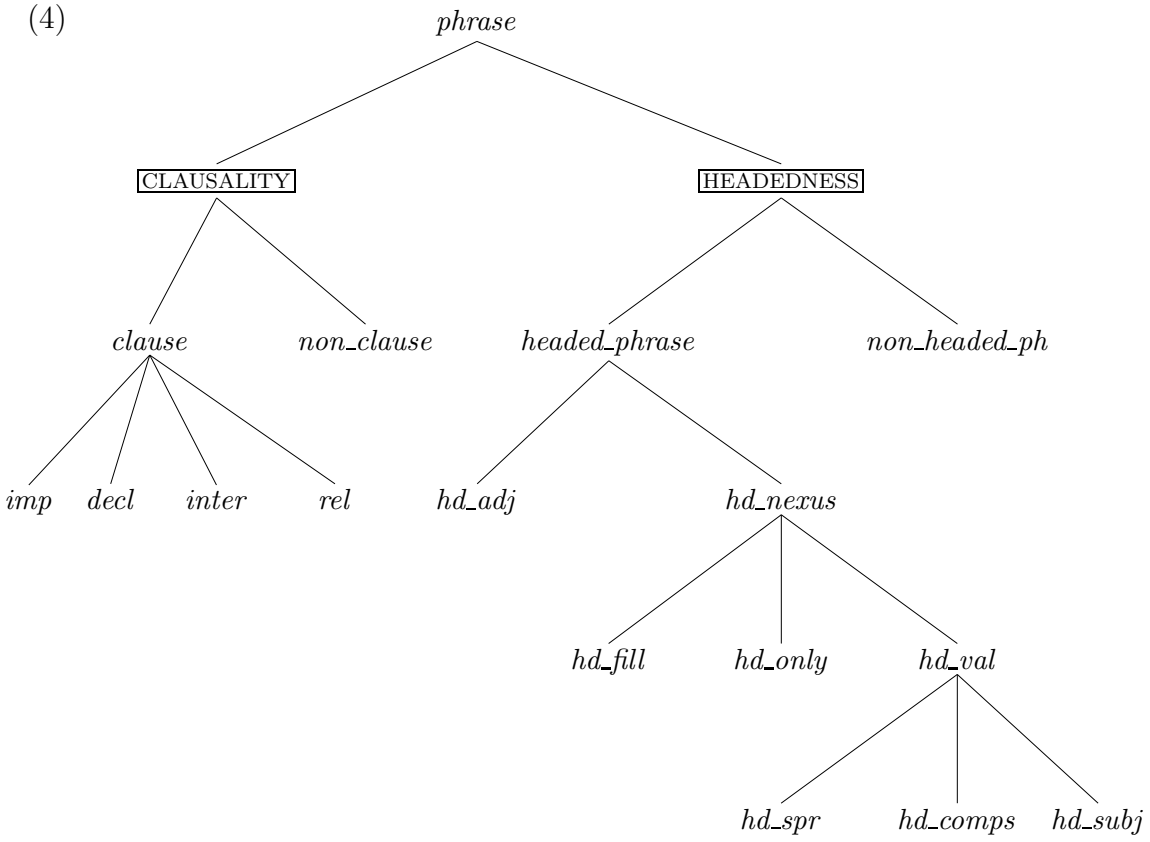
In HPSG the types are partially ordered by subsumption. This means that the feature structures are organized into a multiple inheritance hierarchy, where higher types (also called supertypes) express broader generalizations, and lower types (also called subtypes) contain more specific details. A simple example of such a hierarchy is given in (3). Properties that are shared by all *word*s, *verbs*, *transitive_verbs*, etc., have to be listed only once, and are inherited by their subtypes.

(3)



Sag (1997) and Ginzburg and Sag (2000) show how cross-classifying phrasal hierarchies can be used to account for relative and interrogative clauses in English without any need to stipulate empty elements. Phrases are classified according to two dimensions, CLAUSALITY and HEADEDNESS, as can be seen in the hierarchy in (4). Phrases can be either *clause*s or *non-clause*s, and there are different types of *clause*s—*imperatives*, *declaratives*, *interrogatives*, and *relatives*. *Headed-phrase*s are classified into *head-complement-phrase*s, *head-subject-phrase*s, etc., according to the kind of subcategorization requirement they discharge. There are various constraints associated with these types, specifying for example how the meaning of a phrase is determined from the meanings of its parts (Semantics Principle) and how information percolates up from syntactic heads (Head Feature Principle), as is discussed in more detail in Section 1.2.5. This system of hierarchical classification, which is already in place, can be used for idioms as well.

(4)



1.2.3 The Lexicon

This dissertation treats lexical entries as types, and makes crucial use of the hierarchical structure of the lexicon which is supported by HPSG. This idea was first developed in Flickinger et al. (1985), in Chapter 8 of Pollard and Sag (1987), and in more detail in Flickinger (1987). In these approaches, inflectional and derivational morphology is assumed to be dealt with by lexical rules.

The status and interpretation of lexical rules has been discussed extensively in the literature. Copestake (1992) has integrated them into the hierarchical structure of the lexicon, so that subsumption relationships between lexical rules can be expressed. Krieger and Nerbonne (1993), Riehemann (1993), Koenig and Jurafsky (1994), and Kathol (1998) have proposed various ways of eliminating the need for a separate lexical rule mechanism and expressing these relationships within the feature structure formalism.

1.2.4 MRS

The semantic information is expressed in MRS (Minimal Recursion Semantics), as developed in CSLI's Linguistic Grammars Online (LinGO) project and described in Copestake et al. (1995) and Copestake et al. (1999). Most semantic information in MRS is contained under the feature LZT,[4] which takes a list of *rel*s (relations) as its value. Verb *rel*s have a feature EVENT, which takes a Davidsonian event variable, and *index*-valued features such as ACT(or) and UND(ergoer) (Davis 1996, Davis 2001).[5] Common nouns have *rel*s with the feature INST, which takes an *index* as its value. Each *rel* also has a feature HNDL, which is used as an identifying label to simulate quantifier scoping using ARG(ument) values. The Semantics Principle states that the LZT of a phrase is formed by appending the LZTs of its daughters.

As in the LinGO project I use a distinct *rel* for each meaning of a word, and for each 'synonymous' word, under the assumption that there are few or no exact synonyms. This means that the two lexical entries for ambiguous words like *bank* do not both contain the same *_bank_rel*. It also means that *snake* would have a *_snake_rel* in its semantics and *serpent* a *_serpent_rel*.[6] So the *rel*s can be used to uniquely identify words occurring in idioms, in which (near) synonyms cannot usually be exchanged (*beat around the bush*, #*beat around the shrub*).[7] The leading underscore e.g. in *_eggplant_rel* is used as a notation to mark the types at the bottom of the *rel* hierarchy which correspond to specific lexical items. Idiomatic senses of a word are indicated by *i_*, for example, the *i_bean_rel* of the word *bean* in *spill the beans* is not the same as *_bean_rel* but it is also not the same as *_secret_rel*. I use this notation because the meaning of many idiomatic words does not correspond exactly to the meaning of another word of English.

[4] In earlier versions of MRS this feature was called LISZT and the feature HNDL was called HANDEL, because MRS is a 'compositional semantics'. This was changed because of the distress it caused several readers.

[5] MRS instead uses features like ARG1 for ACT and ARG3 for UND.

[6] The relationship between them could be expressed at a higher level in the hierarchy of *rel*s.

[7] In this dissertation I use #to mark examples that do not have an idiomatic interpretation.

1.2.5 An example

In order to to analyze the sentence in (5) in HPSG, we need lexical entries for *Diana*, *spill*, *the*, and *water*, and phrasal types expressing how these can be combined. These will be considered in turn.

(5) Diana spilled the water.

The sign for the past tense form *spilled* is given in (6). Note that not all this information has to be specified in the lexical entry, as much of it is shared by other transitive verbs and can be inherited by a supertype. Further information has been added by inflectional processes.

(6)
$$\left[\begin{array}{l} \textit{spill} \ \& \ \textit{trans\_verb} \\ \textsc{phon} \ \langle \text{spilled} \rangle \\ \textsc{synsem} \left[\begin{array}{l} \textsc{cat} \left[\begin{array}{l} \textsc{head} \ \textit{verb} \\ \textsc{subj} \left\langle \text{NP}_{\boxed{2}} \right\rangle \\ \textsc{comps} \left\langle \text{NP}_{\boxed{3}} \right\rangle \end{array}\right] \\ \textsc{cont} \left[\begin{array}{l} \textit{mrs} \\ \textsc{index} \ \boxed{1} \ [\textsc{tense} \ \textit{past}] \\ \textsc{lzt} \left\langle \left[\begin{array}{l} \textit{\_spill\_rel} \\ \textsc{event} \ \boxed{1} \\ \textsc{act} \ \boxed{2} \\ \textsc{und} \ \boxed{3} \end{array}\right] \right\rangle \end{array}\right] \end{array}\right] \end{array}\right]$$

The verb *spilled* specifies that it wants to combine with an NP COMP(lement) and an NP SUBJ(ect). Its semantics is a *_spill_rel*(ation), which is an EVENT with an ACT(or) and an UND(ergoer), that correspond to the subject and the complement, respectively. Note that in the notation $\text{NP}_{\boxed{1}}$ the boxed number $\boxed{1}$ is the semantic INDEX of the NP, i.e. $\text{NP}_{\boxed{1}}$ is an abbreviation for (7).

(7)
$$\begin{bmatrix} & \begin{bmatrix} \text{HEAD } \textit{noun} \\ \text{SPR } < \quad > \\ \text{COMPS } < \quad > \end{bmatrix} \\ \text{CAT} & \\ \text{CONT} \mid \text{INDEX } \boxed{1} & \end{bmatrix}$$

The AVMs in this example have also been simplified in other ways to make the exposition more manageable. For example, I have left out the handels (HNDL), argument structure (ARG-ST) and the LOCAL and NON-LOCAL features because I am not talking about scoping, unbounded dependencies, binding theory, etc., in this example. I also abbreviated the AVMs in other ways. For example, I did not indicate general types like *synsem* which do not add any content. The types that are given are shown in italics in the top left corner of a pair of square brackets. When two types are given with an & symbol this means that the AVM is of both types, either because it inherits from both of them or because one of the types is a supertype of the other. This information is not usually given in AVMs because it can be inferred from the type hierarchy, but I include it here for expository purposes. Note that throughout this dissertation I treat lexical entries as types, e.g. *spill* above is a subtype of *trans_verb*. Angle brackets indicate lists, and $<\ >$ stands for the empty list. The elements of lists are also of a particular type, which is not shown in the AVMs. The elements of the valence feature lists like SUBJ and COMPS are of type *synsem*, and the elements of the semantic LZT are *rel*s.

The information for the name *Diana* is shown in (8). Again much of this information is the same for all names and does not have to be listed separately for each name.

(8)
$$\begin{bmatrix} \textit{diana} \ \& \ \textit{name} \\ \textsc{phon} \ \langle \text{diana} \rangle \\ \textsc{synsem} \begin{bmatrix} \textsc{cat} \begin{bmatrix} \textsc{head} \ \textit{noun} \\ \textsc{spr} \ \langle \ \rangle \\ \textsc{comps} \ \langle \ \rangle \end{bmatrix} \\ \textsc{cont} \begin{bmatrix} \textsc{index} \ \boxed{1} \\ \textsc{lzt} \left\langle \begin{bmatrix} \textit{\_diana\_rel} \\ \textsc{inst} \ \boxed{1} \end{bmatrix} \right\rangle \end{bmatrix} \end{bmatrix} \end{bmatrix}$$

The noun *water* is described as in (9). Note that nouns equate their INST and INDEX values. Again, other *mass_noun*s have many of the same properties. For example, all such nouns take an optional specifier (SPR), indicated by the parentheses.

(9)
$$\begin{bmatrix} \textit{water} \ \& \ \textit{mass\_noun} \\ \textsc{phon} \ \langle \text{water} \rangle \\ \textsc{synsem} \begin{bmatrix} \textsc{cat} \begin{bmatrix} \textsc{head} \ \textit{noun} \\ \textsc{spr} \ \langle ([\ \textsc{synsem} \,|\, \textsc{cat} \,|\, \textsc{head} \ \textit{det} \ ]) \rangle \\ \textsc{comps} \ \langle \ \rangle \end{bmatrix} \\ \textsc{cont} \begin{bmatrix} \textsc{index} \ \boxed{1} \\ \textsc{lzt} \left\langle \begin{bmatrix} \textit{\_water\_rel} \\ \textsc{inst} \ \boxed{1} \end{bmatrix} \right\rangle \end{bmatrix} \end{bmatrix} \end{bmatrix}$$

The description for the definite determiner *the* is given in (10). It takes the INDEX of the noun it specifies (SPEC) as a value of its bound variable (BV).

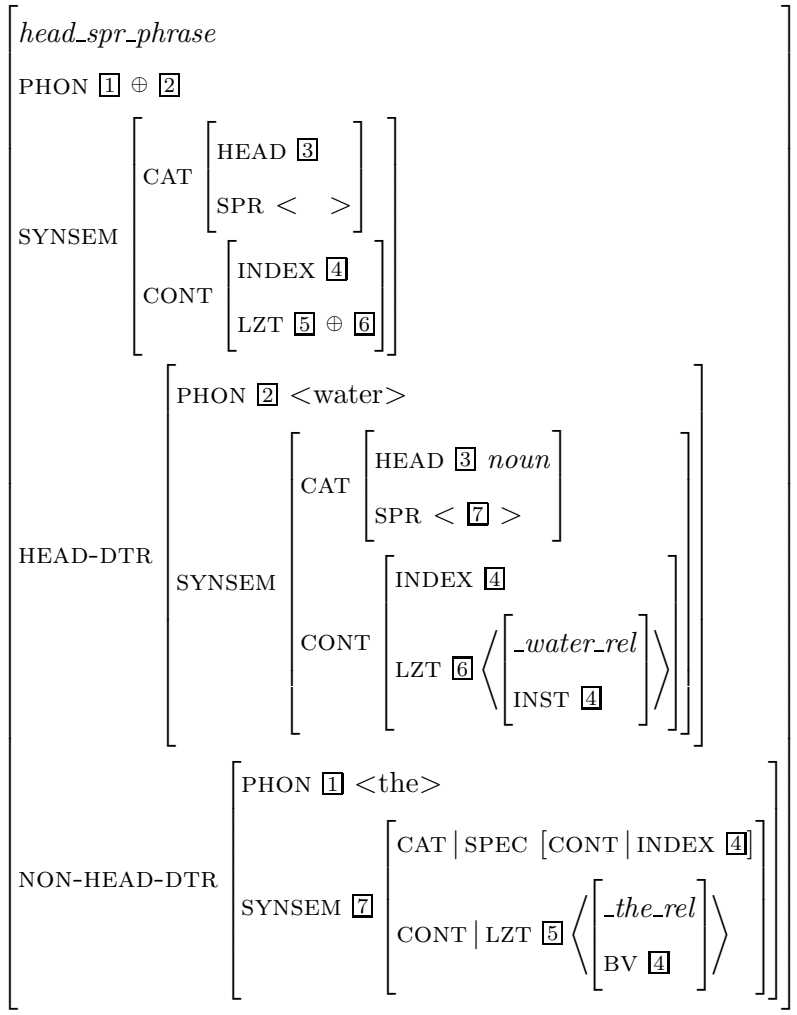
(10)
$$\begin{bmatrix} \textit{the} \ \& \ \textit{determiner} \\ \textsc{phon} \ \langle \text{the} \rangle \\ \textsc{synsem} \begin{bmatrix} \textsc{cat} \begin{bmatrix} \textsc{head} \ \textit{det} \\ \textsc{spr} \ \langle ([\ ]) \rangle \\ \textsc{spec} \ [\textsc{cont} \mid \textsc{index} \ \boxed{1}] \end{bmatrix} \\ \textsc{cont} \begin{bmatrix} \textit{mrs} \\ \textsc{index} \ \boxed{1} \\ \textsc{lzt} \left\langle \begin{bmatrix} \textit{\_the\_rel} \\ \textsc{bv} \ \boxed{1} \end{bmatrix} \right\rangle \end{bmatrix} \end{bmatrix} \end{bmatrix}$$

These lexical entries combine with the help of phrasal constructions, which are also AVMs that encode constituent structure using daughter features like HEAD-DTR and NON-HEAD-DTR. The determiner *the* and the noun *water* can combine using the *head_specifier_phrase* (abbreviated *head_spr_phrase*) shown in (11), because the description for *water* unifies with the constraints given for the the HEAD-DTR of that phrase, and the description for *the* unifies with the constraints on the NON-HEAD-DTR. The constraints on the SPR value of the noun are 'enforced' because it is structure shared with the SYNSEM of the NON-HEAD-DTR, i.e. the determiner, so the information has to unify.[8] Note that the SPR value of the mother is empty, because this requirement has been satisfied. The LZT of the whole phrase is the result of appending the LZTs of the daughters, which is indicated by the $\oplus$ symbol. For simplicity the information the values of the PHON(ology) are simply appended, and standard orthography is used instead of actual phonological representations.

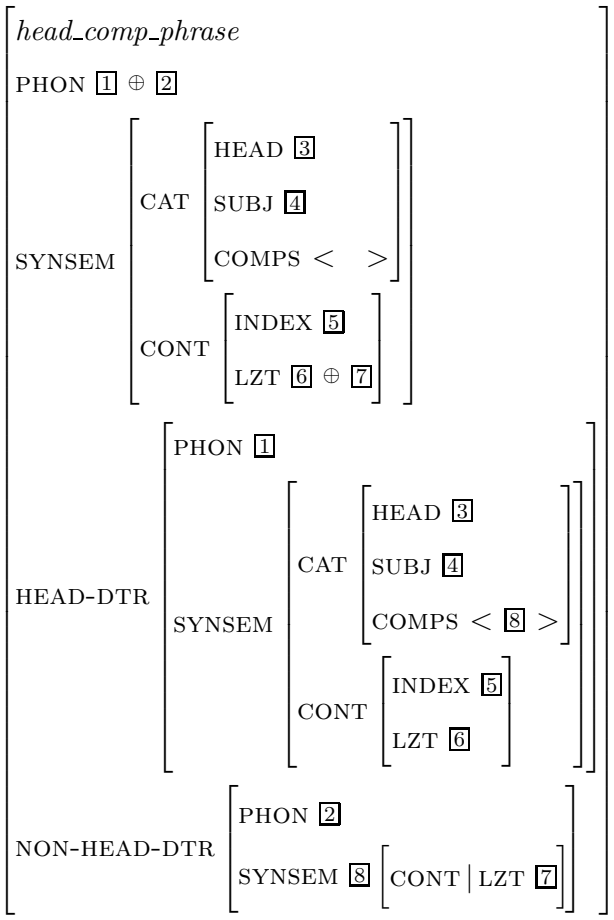
[8]Doing the same thing with the SPEC value would result in a cyclic structure, which is a problem for some formalisms. In any case, it is sufficient to ensure that the INDEX information is shared.

(11)
$$\begin{bmatrix} \textit{head\_spr\_phrase} \\ \text{PHON } \boxed{1} \oplus \boxed{2} \\ \text{SYNSEM} \begin{bmatrix} \text{CAT} \begin{bmatrix} \text{HEAD } \boxed{3} \\ \text{SPR } < \quad > \end{bmatrix} \\ \text{CONT} \begin{bmatrix} \text{INDEX } \boxed{4} \\ \text{LZT } \boxed{5} \oplus \boxed{6} \end{bmatrix} \end{bmatrix} \\ \text{HEAD-DTR} \begin{bmatrix} \text{PHON } \boxed{2} \\ \text{SYNSEM} \begin{bmatrix} \text{CAT} \begin{bmatrix} \text{HEAD } \boxed{3} \\ \text{SPR } < \boxed{7} > \end{bmatrix} \\ \text{CONT} \begin{bmatrix} \text{INDEX } \boxed{4} \\ \text{LZT } \boxed{6} \end{bmatrix} \end{bmatrix} \end{bmatrix} \\ \text{NON-HEAD-DTR} \begin{bmatrix} \text{PHON } \boxed{1} \\ \text{SYNSEM } \boxed{7} \begin{bmatrix} \text{CAT} \mid \text{SPEC } \langle [\text{CONT} \mid \text{INDEX } \boxed{4}] \rangle \\ \text{CONT} \mid \text{LZT } \boxed{5} \end{bmatrix} \end{bmatrix} \end{bmatrix}$$

The result of unifying *water* and *the* with the daughters of the *head_spr_phrase* is shown in (12). Because of structure sharing the PHON, HEAD, INDEX, and LZT information from the daughters is now present at the level of the whole phrase as well.

(12)


Verbs combine with their complements using the *head_complement_phrase*, abbreviated *head_comp_phrase*. For simplicity the version I give here accommodates only one complement, although the more general version in Sag (1997) allows for multiple complements. The COMPS subcategorization requirement of the HEAD-DTR is satisfied in the same way as the SPR requirement, by making sure it corresponds to the SYNSEM of the NON-HEAD-DTR using structure sharing. Because at the level of the whole phrase the requirement has been satisfied, the COMPS value of the mother is empty. The SUBJ requirement is not yet satisfied, so it is passed up to the mother. Again the LZT of the whole phrase is the result of appending the LZTs of the daughters.

(13)


Like with lexical entries, the information for these phrasal types can be partially inherited from more general types. For example, all *headed_phrase*s specify that the HEAD and INDEX values are shared between the mother and the head-daughter. And for all *phrase*s the LZT of the mother is the result of appending the LZTs of the daughters, plus any constructional meaning contained in the feature C-CONT, as shown in (14).[9]

[9]The reader may have noticed that I appended the LZTs in the other order in the constraint on *head_spr_phrase*, so that the order of LZT elements corresponds to the normal English word order. But because the order of elements in the LZT is not important it would be no problem to have a general constraint as in (14). If a different order is desired to make the result more human-readable, the constraint can be stated separately on a *head-initial* and *head-final* subtype.

(14)
$$\left[\begin{array}{l} \textit{headed\_phrase} \\ \textsc{synsem}\,|\,\textsc{cont}\,|\,\textsc{lzt}\ \boxed{1} \oplus \boxed{2} \oplus \boxed{3} \\ \textsc{c-cont}\ \boxed{3} \\ \textsc{head-dtr}\,|\,\textsc{synsem}\,|\,\textsc{cont}\,|\,\textsc{lzt}\ \boxed{1} \\ \textsc{non-head-dtr}\,|\,\textsc{synsem}\,|\,\textsc{cont}\,|\,\textsc{lzt}\ \boxed{2} \end{array}\right]$$

Note that the constraint in (14) is sufficient only for binary branching grammars. To be more general, the constraint would look like (15).

(15)
$$\left[\begin{array}{l} \textit{headed\_phrase} \\ \textsc{synsem}\,|\,\textsc{cont}\,|\,\textsc{lzt}\ \boxed{1} \oplus \boxed{2} \oplus \boxed{3} \oplus \ldots \boxed{n} \\ \textsc{c-cont}\ \boxed{1} \\ \textsc{head-dtr}\,|\,\textsc{synsem}\,|\,\textsc{cont}\,|\,\textsc{lzt}\ \boxed{2} \\ \textsc{non-head-dtrs}\left\langle [\ldots\textsc{lzt}\ \boxed{3}],\ \ldots,\ [\ldots\textsc{lzt}\ \boxed{n}] \right\rangle \end{array}\right]$$

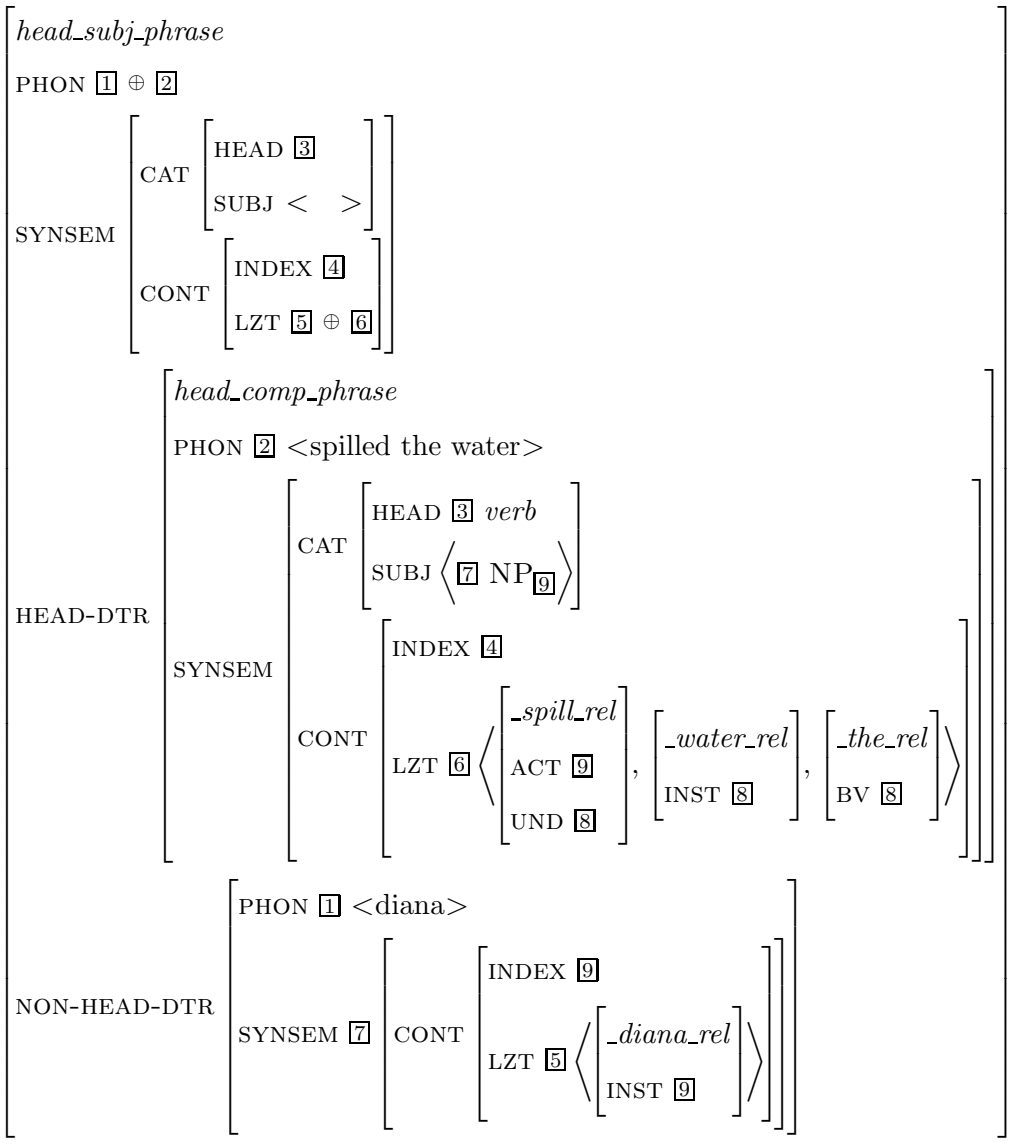
The result of unifying *spilled* with the HEAD-DTR of the *head_comp_phrase* and *the water* with the NON-HEAD-DTR is shown in (16). Note that for space reasons the DTRs features of the *head_spr_phrase* are not shown.

(16)
$$\left[\begin{array}{l} \textit{head\_comp\_phrase} \\ \textsc{phon}\ \boxed{1} \oplus \boxed{2} \\ \textsc{synsem}\left[\begin{array}{l} \textsc{cat}\left[\begin{array}{l} \textsc{head}\ \boxed{3} \\ \textsc{subj}\ \boxed{4} \\ \textsc{comps}\ <\quad> \end{array}\right] \\ \textsc{cont}\left[\begin{array}{l} \textsc{index}\ \boxed{5} \\ \textsc{lzt}\ \boxed{6} \oplus \boxed{7} \end{array}\right] \end{array}\right] \\ \textsc{head-dtr}\left[\begin{array}{l} \textsc{phon}\ \boxed{1}\ \text{<spilled>} \\ \textsc{synsem}\left[\begin{array}{l} \textsc{cat}\left[\begin{array}{l} \textsc{head}\ \boxed{3}\ \textit{verb} \\ \textsc{subj}\ \boxed{4} \\ \textsc{comps}\left\langle \boxed{8}\ \text{NP}_{\boxed{9}} \right\rangle \end{array}\right] \\ \textsc{cont}\left[\begin{array}{l} \textsc{index}\ \boxed{5} \\ \textsc{lzt}\ \boxed{6}\left\langle \left[\begin{array}{l} \textit{\_spill\_rel} \\ \textsc{und}\ \boxed{9} \end{array}\right] \right\rangle \end{array}\right] \end{array}\right] \end{array}\right] \\ \textsc{non-head-dtr}\left[\begin{array}{l} \textit{head\_spr\_phrase} \\ \textsc{phon}\ \boxed{2}\ \text{<the water>} \\ \textsc{synsem}\ \boxed{8}\left[\textsc{cont}\left[\begin{array}{l} \textsc{index}\ \boxed{9} \\ \textsc{lzt}\ \boxed{7}\left\langle \left[\begin{array}{l} \textit{\_water\_rel} \\ \textsc{inst}\ \boxed{9} \end{array}\right], \left[\begin{array}{l} \textit{\_the\_rel} \\ \textsc{bv}\ \boxed{9} \end{array}\right] \right\rangle \end{array}\right]\right] \end{array}\right] \end{array}\right]$$

Verb phrases combine with their subjects using the *head_subj_phrase*, abbreviated as *head_subj_phrase*, shown in (17). Again the LZT of the whole phrase is the result of appending the LZTs of the daughters. Also note that subject-verb agreement is enforced because the SUBJ value of the HEAD-DTR is structure shared with the SYNSEM of the NON-HEAD-DTR. For example a 3rd person singular verb form like *spills*, which wants to combine with an NP[*3sg*] will not be able to combine with subjects like *we*, with properties that conflict with this requirement so that unification fails.

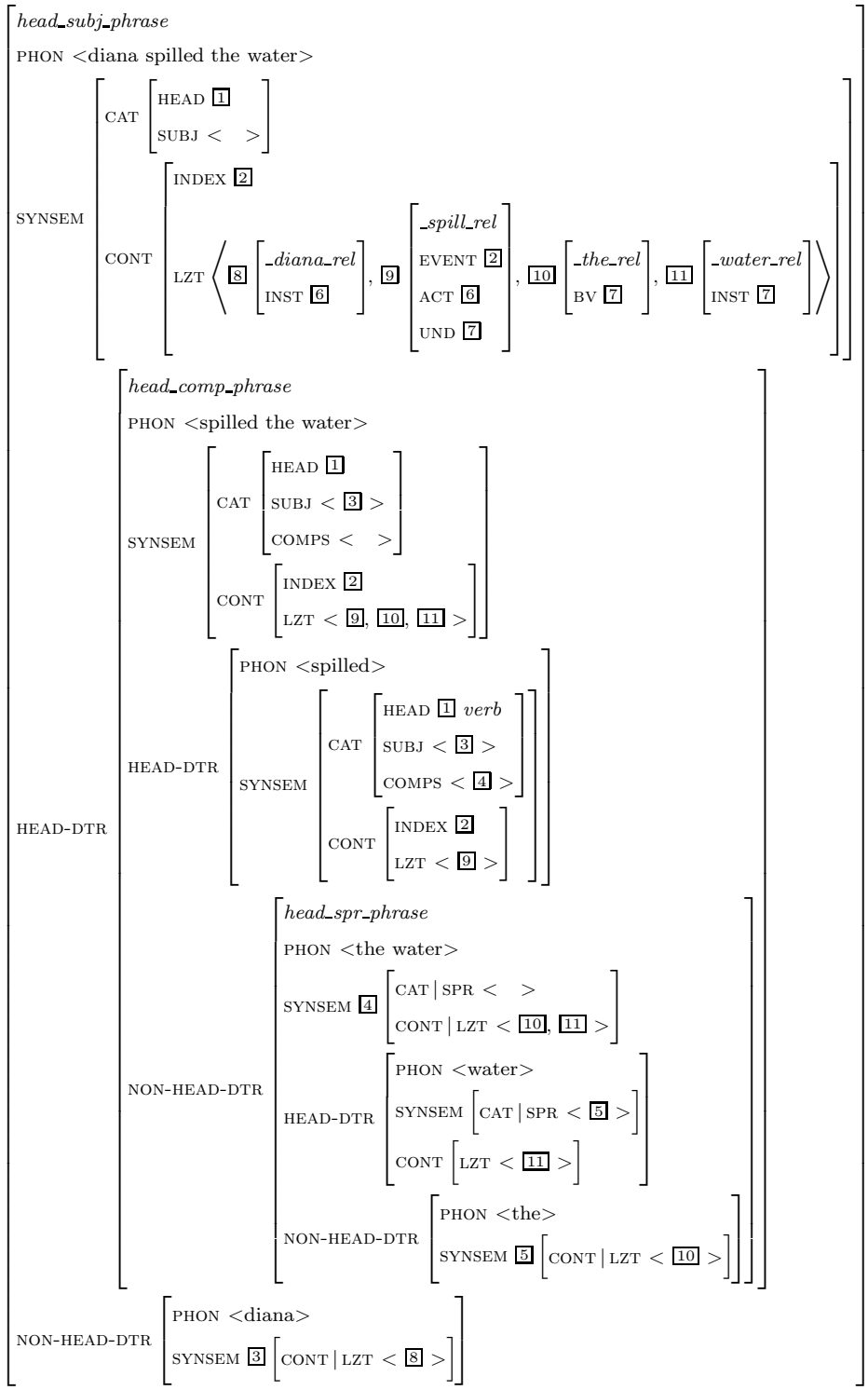
(17)
$$\left[\begin{array}{l} \textit{head\_subj\_phrase} \\ \text{PHON } \boxed{1} \oplus \boxed{2} \\ \text{SYNSEM} \left[\begin{array}{ll} \text{CAT} & \left[\begin{array}{l} \text{HEAD } \boxed{3} \\ \text{SUBJ } < \quad > \end{array}\right] \\ \text{CONT} & \left[\begin{array}{l} \text{INDEX } \boxed{4} \\ \text{LZT } \boxed{5} \oplus \boxed{6} \end{array}\right] \end{array}\right] \\ \text{HEAD-DTR} \left[\begin{array}{ll} \text{PHON } \boxed{2} & \\ \text{SYNSEM} & \left[\begin{array}{ll} \text{CAT} & \left[\begin{array}{l} \text{HEAD } \boxed{3}\ \textit{fin} \\ \text{SUBJ } < \boxed{7} > \end{array}\right] \\ \text{CONT} & \left[\begin{array}{l} \text{INDEX } \boxed{4} \\ \text{LZT } \boxed{6} \end{array}\right] \end{array}\right] \end{array}\right] \\ \text{NON-HEAD-DTR} \left[\begin{array}{ll} \text{PHON } \boxed{1} & \\ \text{SYNSEM } \boxed{7} & \left[\text{CONT} \mid \text{LZT } \boxed{5}\right] \end{array}\right] \end{array}\right]$$

The result of unifying *Diana* with the NON-HEAD-DTR and *spilled the water* with the HEAD-DTR of this phrase results in the description in (18).

(18)


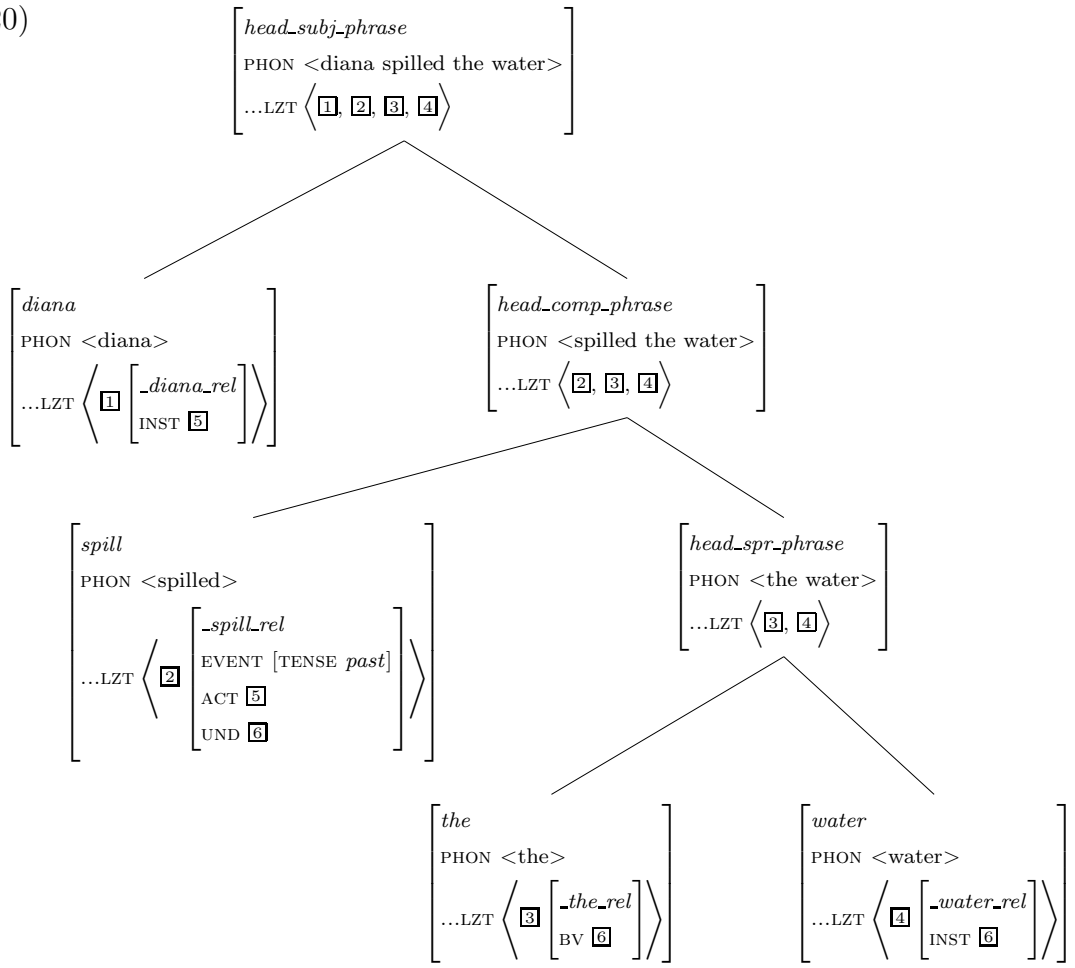
This description is again abbreviated in that it leaves out the daughters' daughters. A somewhat more complete version including the whole 'syntax tree' is shown in (19). Note that in (19), unlike in (18), I show the results of the append operation for the PHON(ology) instead of showing all the instances of appending using structure sharing, to avoid having a confusingly large number of tags in the representation, and to make it easier to see where each phrase is located in the representation.

(19)


head_subj_phrase
PHON <diana spilled the water>
SYNSEM
CAT
HEAD 1
SUBJ < >
CONT
INDEX 2
LZT
8 _diana_rel INST 6
9 _spill_rel EVENT 2 ACT 6 UND 7
10 _the_rel BV 7
11 _water_rel INST 7
HEAD-DTR
head_comp_phrase
PHON <spilled the water>
SYNSEM
CAT
HEAD 1
SUBJ < 3 >
COMPS < >
CONT
INDEX 2
LZT < 9, 10, 11 >
HEAD-DTR
PHON <spilled>
SYNSEM
CAT
HEAD 1 verb
SUBJ < 3 >
COMPS < 4 >
CONT
INDEX 2
LZT < 9 >
NON-HEAD-DTR
head_spr_phrase
PHON <the water>
SYNSEM 4
CAT | SPR < >
CONT | LZT < 10, 11 >
HEAD-DTR
PHON <water>
SYNSEM CAT | SPR < 5 >
CONT LZT < 11 >
NON-HEAD-DTR
PHON <the>
SYNSEM 5 CONT | LZT < 10 >
NON-HEAD-DTR
PHON <diana>
SYNSEM 3 CONT | LZT < 8 >

In a more familiar tree notation this example would look like (20). The ... FEATURE notation indicates that the full path leading to the FEATURE is not given, so as to save space.

(20)

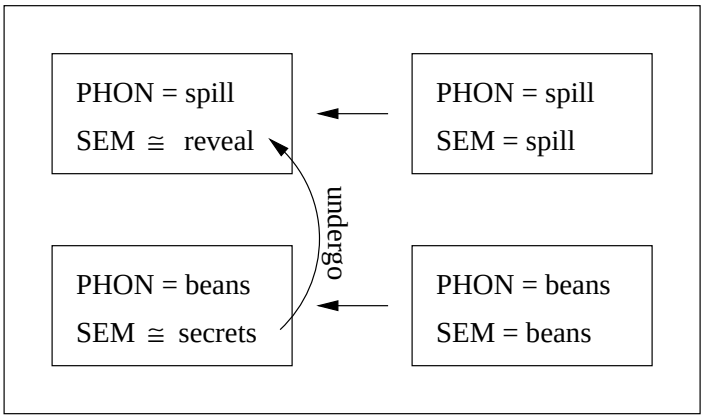


Note that while I showed in a particular order how the AVM in (19) is assembled, this is irrelevant in the formalism, as the set of constraints in an HPSG grammar is declarative and order-independent. All that matters is that none of the constraints are violated and that the complete utterance is of a type that can be a *root_phrase*.

1.3 Preview of the Constructional Approach

The analysis for an idiomatic sentence like *Diana spilled the beans* will look very much like that in (19) and (20). One obvious difference is that the words *spill* and *beans* will have their idiomatic meanings, *i_spill_rel* and *i_bean_rel*. But the syntactic structure of the sentence is the same, and the same phrasal construction types, like *head_comp_phrase* etc., are used to license it. The main difference is that there are co-occurrence restrictions between *spill* and *beans*, which are not expressed as sub-categorization constraints using valence features like COMPS. Instead, a mechanism will be presented that allows specifying simply that this idiomatic phrase contains the two idiomatic words *spill* and *beans* standing in the right semantic relationship to each other. This idiomatic phrase will act as a 'lexical entry' for the whole idiom, so that separate lexical entries for the idiomatic words are not needed. The relationship of these idiomatic words to the corresponding literal lexical entries is also expressed. An intuitive depiction of this approach can be seen in (21).

(21)



The outer box corresponds to the whole phrase, which contains the words *spill* and *beans* with their idiomatic meaning. The right-to-left arrows indicate the relationship of these words to the literal lexical entries for *spill* and *beans*, which expresses the metaphorical relationship and accounts for the fact that many properties are shared. How this works formally will be shown in detail in Chapter 5.

Chapter 2

Data that Approaches to Idioms Need to Capture

2.1 Introduction

This chapter gives data that approaches to idioms need to capture, and argues that these data require an approach to idioms that is semantic and constructional. First I discuss why idioms need to be represented phrasally. The main evidence for this is that idiomatic words are not free, that there are canonical forms of idioms which need to be represented, that there has to be a locus for the metaphorical mapping and for the semantics of non-decomposable idioms, and that many idioms involve more than just head-argument relationships of the type expressed via valence constraints.

The main reason that previous approaches have not been phrasal is the problem of variation. Several types of variation are distinguished and I mention briefly what problems they pose to standard constructional, as well as some word-level approaches. This is discussed in more detail in Chapter 4. Some kinds of variation involve semantic modification of idiom parts, which shows that the parts of some idioms have to correspond to parts of the idiomatic meaning. This contrasts with claims in the literature (Katz 1973, Chomsky 1980) that idioms cannot be given a compositional analysis, and confirms the findings of Nunberg et al. (1994).

Much of the data in this chapter is taken from the corpus study in Chapter 3. This

is indicated for each example or set of examples for which it is the case. The remaining data and the unacceptable examples are from the literature or were constructed to illustrate a point for which there were no corpus data available.

2.2 Need for Phrasal Pattern

2.2.1 Absence of Literal Meaning

The literal meanings of many words that are parts of idioms play no role in the meaning of the idiom. For example, there is no *_bean_rel* in the meaning of *spill the beans*, i.e. the meaning of this idiom does not have a component that is a kind of legume. This may sound obvious, but some approaches cannot handle it adequately. For approaches relying on subcategorizing for the phonology, and for phrasal approaches specifying the phonology, it is the non-idiomatic lexical entry for the words which is being selected for/included, and there is no adequate mechanism for discarding the meaning of that entry. For approaches subcategorizing for the syntax or the semantics and for the remaining phrasal approaches, the same problem potentially holds, but can be solved by subcategorizing for/including not the ordinary lexical entries but special idiomatic ones. However, this leads to the problem of having to prevent the special idiomatic entries from occurring by themselves. This is discussed in more detail in Chapter 4.

2.2.2 Idiomatic Words Are Not Free

I define 'idiomatic words' as words that do not exist as independent words with the same meaning, so not all words that occur in idioms are idiomatic words in that sense. But words that meet my definition of 'idiomatic word' can only have one of their literal meanings when they occur by themselves. For example, it is not possible to use *beans* with the meaning it has in *spill the beans*, i.e. roughly 'secrets', when *spill* is not present:

(22) #I would like to divulge some beans about the President.

Furthermore, some idioms like *by dint of* include words like *dint* that never occur outside the idiom. Note that it follows from this that they never occur with this meaning except as part of this larger phrase, which means they fall under my definition of idiom. These words clearly should be represented as part of a phrasal pattern, and this pattern cannot be a simple fixed phrase because 10% of the occurrences of *by dint of* show variations like (23):

(23) a. And people who want to improve their memory could take a few tips from those who, either by virtue of their being "naturally" more creative or **by pure dint of** effort, manage their responsibilities with elephantine aplomb.
b. This month's headlines have had a chilling effect on innocent citizens who **by mere dint of** their ethnicity have become the drive-by victims of the two parties' election-year crossfire.
c. [...] got his position **by sheer dint of** bringing in a lot of money [...]
(from the North American News Corpus)

The same is true for those speakers for whom *hooky* is not an independently existing word, yet variations like in (24) are still possible.

(24) To the chagrin of his speech-writing team, the president has **played a bit of hooky** during his rail trip to Chicago, as he took time to mingle with admirers and hail onlookers from the rear platform of the train.
(from the North American News Corpus)

In an approach subcategorizing for an idiomatic complement for which there is a separate lexical entry, this fact that idiomatic words are not 'free' is not captured—only one of the idiomatic words is constrained to select for the other, but the reverse is not true. For example, the idiomatic verb *spill* can say that it only takes *beans* as a complement, but there is no way of saying in a lexical entry for the idiomatic noun *beans* that it only occurs as the complement of *spill*.

So a whole separate mechanism is needed for controlling the distribution of idiomatic words like *beans*,[1] preventing them from having their idiomatic meaning when

[1]For some non-decomposable idioms like *kick the bucket* one actually has to worry about a whole idiomatic NP like *the bucket*. Canonical forms for idioms like *spill the beans* also involve fixed

they do not occur as part of the idiom. I have not seen such a mechanism proposed in the literature. The problem is compounded by the fact that the meaning of idiomatic *beans* cannot be made incompatible with that of other verbs in general, because it can occur with them as long as idiomatic *spill* is also present in the right semantic relationship to it. A corpus example for this involving the idiom *pull strings* can be seen in (25)—further examples can be found in Section 2.3.6. In this example *strings* is 'licensed' because *pulled* occurs in the relative clause, so that *strings* is still the UNDERGOER of *pull*. However, *strings* is also the complement of *justify*.

(25) [...] Robert McNamara's new book **justified all the strings Clinton pulled** as a young man [...] (from the North American News Corpus)

This problem is even worse for idioms that involve more than two fixed items—each of them would need to be constrained to occur with all of the others. Note also that even if a technical solution to this problem could be found, there is no motivation for attributing the idiomaticity of the expression solely to the verb. On the contrary—for many idioms, the verb can be used in its idiomatic, metaphorical meaning outside of the idiom, at least by some speakers. For example, a corpus search found 63 examples of *spill secrets* and 6 examples of *spill information*. In contrast, there are no corpus examples of #*reveal (the) beans*, #*divulge (the) beans*, or #*pour out (the) beans*. Apparently some speakers have an independent lexical entry for *spill* with the same meaning it has in the idiom, i.e. for them the only thing that makes *spill the beans* idiomatic is the unusual meaning for *beans*. Note that for such speakers, *spill* is not what I call an 'idiomatic word', although of course it is still a word that is part of an idiom. This is because I define 'idiomatic word' as a word that does not have an existence as an independent lexical entry with the same meaning.

This state of affairs is exactly the opposite from what is expressed in an approach that represents *spill the beans* as an unusual lexical entry for *spill* that only occurs with the complement *the beans*, referring to an independent lexical entry for *beans* with the meaning *secrets*.

specifiers, but those have their usual meanings, i.e. they are not 'idiomatic words'. Instead the canonical form of *spill the beans* includes the normal lexical entry for *the*.

This problem is even more obvious for idioms in which the verb has its literal meaning, e.g. *miss the boat*. In a word-level approach there has to be a special lexical entry for *miss* that occurs only with *boat*, in spite of the fact that in this idiom the verb clearly should not be 'blamed' for the idiomaticity.

In a constructional approach the fact that idiomatic words are not free is captured straightforwardly.[2] There simply is no independent lexical entry for these words in their idiomatic meaning, so they have that meaning only when they are part of the idiomatic phrase.

2.2.3 No Literal Parse

Some idioms are syntactically ill-formed in the sense that they cannot be parsed with the standard lexical entries.

(26) a. trip the light fantastic[3]
b. by and large
c. close up shop
d. find fault
e. set foot

For example, the noun *foot* in English cannot usually occur in the singular without a determiner as it is not a mass noun. Further examples can be found in Nunberg et al. (1994:515). This makes it even clearer that it is necessary to prevent idiomatic lexical entries from occurring by themselves. In these cases that would not just result in an impossible interpretation, but in ungrammatical sentences being parsed (e.g. **there is much foot in the room*).

[2]As discussed in Chapter 5 this turns out to be a little more complicated than one would think in a formal approach, although the solution given there presupposes a phrasal analysis.

[3]Note that it is not possible to treat this idiom as completely fixed: there is a corpus example of *tripping a bit of the light fantastic*.

2.2.4 Restricted Flexibility of Some Idioms

Some idioms can never occur actively. For example there are 138 corpus examples of the idiom *caught in the middle* which are clearly not active, although it is somewhat unclear whether they are all adjectival passives with a stative interpretation (Wasow 1977) or whether some of them may be actual verbal passives (see Chapter 3). There are no corpus examples of the form *catch/catches/catching NP in the middle*, and there is only one active example in the corpus, which is unacceptable to at least some native speakers:

(27) ? [. . .] Johnson [. . .] did make one comment that in retrospect hinted at a bombshell that exploded over the industry last week and **caught investors in the middle**.

Further passive idioms are *not born yesterday* (see Chapter 3), *cast in stone*,[4] and *taken aback*.[5]

Other idioms, for example most non-decomposable idioms, cannot be passivized. I found passive examples with only 2 of the 15 non-decomposable idioms I studied, and they are unacceptable to at least some native speakers:

(28) a. ?It was good **the air was cleared**.

b. ??But it increases the odds that if **the COLA bullet is bitten**, retired feds won't be the only group with tooth-marks on their hides.

Nunberg et al. (1994:516) give examples of idioms that occur only in other constructions. For example, *who/what the hell* can only occur preposed, *what's eating NP* can only be a *wh*-question, *Is the Pope Catholic?* has to involve subject-aux inversion, *play hard to get* has to involve tough movement, and *Break a leg!* has to be imperative.

[4]There is actually one active example (out of 22 examples) of this idiom in the corpus: *. . . the Motor Voter Registration law . . . will "cast this phony registration information in stone" . . .*

[5]Only 22 out of 624 occurrences of this idiom (4%) are active, and it seems to be restricted to passive for at least some speakers.

2.2.5 Canonical Forms

As will be shown in detail in Chapter 3, idioms have canonical forms. By that I mean that for each idiom there is a particular fixed phrase (modulo inflection of the head) which is recognized by speakers of the language as the normal form this idioms takes, and which is used much more frequently than would be predicted from independent factors. Examples of canonical forms are *run the show*, *call the shots*, *lose face*, and *make waves*. So the nouns in canonical forms of idioms can be singular or plural, and definite or indefinite, but for each idiom only one of these four possible forms is the canonical one. Some canonical forms have further properties, e.g. *be caught short* is usually passive and *not born yesterday* is usually passive and negative. But other forms of these idioms do occur in the corpus, so they are not impossible for semantic or other reasons. Speakers also know that the canonical form of the idiom *born out of wedlock* is not *born outside of wedlock*, *born outside wedlock*, or *born without benefit of wedlock*, and that the idiom is *save NP for a rainy day* and not *save NP for rainy days* or *save NP for a rainier day*. Here the form of the idiom is clearly determined by convention and not by semantic reasons.

For V+NP idioms the question of what the canonical form is has a straightforward answer—it is the VP, modulo inflection of the head verb, with the NP in the form (for example definite plural) in which it most frequently occurs.[6] For example the canonical form of the idiom *spill the beans* is *spill/spills/spilled/spilling the beans*. For all the V+NP idioms I studied this corresponds to the form in which they are cited in my idioms dictionaries[7] and in the literature.

Studies of paraphrases in Chapter 3 show that there are no such canonical forms for non-idioms like *reveal the secrets*. While the idiom *spill the beans* occurs in this particular form (modulo inflection of *spill*) 87% of the time, its paraphrase *reveal the secrets*, matched to the idiom in number and definiteness, accounts for only 1% of the occurrences of *reveal the/a secret(s)*, and the most frequent form *reveal secrets* accounts for only 7% of the data. Note also that none of the 63 occurrences of *spill*

[6]Things are not quite so clear for some other idioms. This is discussed in Chapter 3.

[7]The dictionaries I used are the Collins COBUILD Dictionary of Idioms and NTC's American Idioms Dictionary. For full citations see Appendix B.

the secrets is of this form (matching the idiom in number and definiteness), and that the most frequent form, *spill secrets*, accounts for only 14% of the examples.

This shows that there is nothing about the meanings of *reveal* and *secret* such that *reveal the secrets* would be the most likely form in which they cooccur. So there does not seem to be a semantic explanation for the definite article *the* or the plural *beans* in *spill the beans*. Further evidence for this is the fact that variations of this idiom with other specifiers are fine. Corpus examples of this include ***some*** *beans*, ***any*** *beans*, ***mountains of*** *beans*, and ***whatever*** *beans*.

One of the reasons for the fact that non-idioms do not have a canonical form is that non-idiomatic nouns are more frequently modified by adjectives. Again, this is not something that is impossible for idioms—modification by adjectives is something that is observed in all the decomposable idioms studied. The best explanation for why it is less frequent with idioms is that speakers have a representation for the canonical form of the idiom, and are therefore more likely to retrieve and use that form unchanged.

I compared the number of NPs containing singular or plural nouns (i.e. excluding pronominal NPs) with the number of NPs containing singular or plural nouns and also at least one adjective in the Treebank corpus. The percentage of NPs containing adjectives was 39% in the Wall Street Journal part of the corpus, 32% in the Brown part, and 24% in the Switchboard part. Of these, the WSJ is obviously the most similar in type to the newspaper corpus I used for my idiom study, as the part of the Brown corpus that is in the Treebank corpus is almost all fiction, and the Switchboard corpus is spoken language.

As can be seen in Table 2.1, the percentage of decomposable idioms from the random sample studied in Chapter 3 in which the idiomatic noun is modified by adjectives is 9%. So non-idiomatic noun phrases are over four times more likely to contain adjectives than idiomatic NPs.

Note that it is not just the average that is well below that of nonidiomatic NPs, but also all individual idioms except for *pay dividends* and *strike a chord*. As will be discussed in Chapter 3, *dividends* can have its figurative meaning in the absence of *pay*, so *pay dividends* is probably a collocation and not an idiom for many speakers.

Decomp. Idioms	**Total # of Tokens**	**# Modified by Adjectives**	**% Modified by Adjectives**
turn the tables	518	11	2%
call the shots	589	15	3%
deliver the goods	176	15	10%
lose face	137	5	4%
make waves	243	22	9%
run the show	368	21	6%
pay dividends	418	165	39%
sound the death knell	110	1	1%
break the mold	168	19	11%
lose ground	2350	59	2%
strike a chord	688	245	36%
rear its head	128	21	16%
break the ice	183	3	2%
level the playing field	443	39	9%
lead the field	169	30	18%
take a back seat	700	13	2%
Total:	7388	684	**9%**

Table 2.1: Percentage of Idioms Modified by Adjectives

This would explain its higher rate of modification. For *strike a chord* the related fixed expression *strike a responsive chord* probably influences the rate of modification—see Chapter 3.

On average the canonical forms of idioms account for about 75% of the occurrences of decomposable idioms and 97% of the occurrences of non-decomposable idioms, and need to be represented in a full account of idioms. Canonical forms lend themselves to a constructional representation because they are by definition more fixed in form. In this dissertation the canonical form of an idiom will be represented as a more fully specified subtype of the constructional representation for that idiom (see Chapter 5).

2.2.6 Idiom Families

Some idioms allow for lexical variation in one of their components, with essentially the same meaning:

(29) a. put/lay/spread your cards on the table
b. set/lay eyes on
c. throw someone to the wolves/lions

Other idioms form families that are closely related semantically:

(30) a. bring/come to light[8]
b. keep/lose one's cool
c. get/start/set/keep/have the ball rolling

Further examples can be found in Nunberg et al. (1994:504). Some of this variation may be due to dialectal differences, but sometimes the different variants have a subtly different meaning, and productive substitution of idiom parts is also possible in some cases. The members of these idiom families should be related in a theory of idioms, and these data suggest that the relationship between idiomatic verbs and their complements is semantic in nature, and that the metaphorical mapping needs to be represented as a whole.

Note that sometimes the variable part is not a word with an idiomatic meaning. For example, the verbs in (30c) can probably have the same meaning in *get/set/start/keep/have the activity going*.[9] However, *get the ball rolling* is still an idiom according to my definition, because a least one of its components cannot have its literal meaning outside of this expression.

Note also that being variable does not necessarily imply that an idiom part is 'free'. For example the words *wolves* and *lions* in the idiom in (29c) are not 'free',

[8]As Binnick (1971) observed, there are many other *bring/come* pairs, such as *bring/come to blows*, *bring/come to a head*, *bring/come forth*, *bring/come down to earth*, etc.

[9]This cannot be determined from the corpus as there are only 2 corpus examples of *get the activity going* and 1 example of *keep the activity going*. For the 155 occurrences of *VERB the ball rolling* the distribution is 52% *get*, 26% *start*, 13% *set*, 8% *keep*, and 1% *have*. This includes the 19 examples in which *ball* is modified, which account for 12% of these occurrences.

because they mean roughly *enemies*, and generally cannot have that meaning outside of the idiom (#*Our company has many wolves.*).[10]

To the extent that the variable words do not belong to an independently existing class (of words, lexemes, or semantic relations, depending on the particular approach) it seems necessary to express the relevant generalization in a supertype (e.g. *large_carnivorous_animal_rel*). The corpus contains 39 instances of this idiom involving *wolves* and 11 instances involving *lions*, and these two variants seem to be conventionalized. However, the metaphor needs to be represented as a higher level generalization as well, to account for corpus occurrences like the following, productively varied forms of this idiom:[11]

(31) a. After that, I'm **thrown to the tigers**.
b. It can't be the case that they can **throw anyone they want to the sharks**
c. There is a whole group of athletes in East Germany that without justification was **thrown to the dogs**

Note that there are probably additional constraints on what 'large carnivorous animals' are likely to be productively substituted in this idiom. I found no corpus examples of *throw NP to the bears/grizzlies/wildcats/panthers/leopards/coyotes/hyenas*. Which of these are more likely to be productively used may depend on factors like how prototypical an example of a dangerous animal they are, how dangerous they are perceived to be to humans, whether they appear in groups, and how similar they are to the conventionalized choices *wolves* and *lions*. Which of these factors are important may vary from speaker to speaker, and it would be hard to formalize them. Some productive examples like *throw NP to the panthers* may be considered word

[10]Because the animal-to-human metaphor is at least semi-productive it may be possible to use words like *wolves* with a similar meaning in a creative metaphor. But presumably such metaphors occur in a more marked context than idioms.

[11]Note that *throw NP to the dogs* is a conventional variant of this idiom for some speakers, and appears in some idioms dictionaries. There is also one occurrence involving *dogs* which has a slightly different meaning ('let NP go to waste'): *South Carolina is throwing its economic-development efforts to the dogs.* This may be a deliberate or accidental case of mixing this idiom with the idiom *go to the dogs*. It is not clear whether (31c) also has this meaning. Note that there is no inanimate complement among the examples involving *wolves* or *lions*—they are all about people being 'abandoned to the enemies'.

play.[12] But expressing an approximate generalization in the grammar is nevertheless desirable, as it helps to make sense of productive substitutions.

Idiom families cannot be accounted for in any approach fixing the phonology. When the variable part is the verbal head of the idiom, a word-level approach would require the variant forms to be represented as lexical entries for all the possible verbs.

2.2.7 Locus for the Metaphorical Mapping

Approaches to idioms differ considerably with respect to their ability to represent the metaphorical mapping between the literal and idiomatic meanings of the expressions involved. In many approaches the connection to the literal meanings is made at the word level, so that only part of the metaphor can be expressed. But it is not sufficient or even correct to say that *beans* is related to *secrets* and that *spill* is related to *reveal*—the real mapping is between *spill the beans* and *reveal the secrets*. In some word-level approaches it may be possible to express that *spill* is related to *reveal* only when its complement is *the beans*, but there is still no way to bring in both the literal and the idiomatic meaning of *beans*, and an independent mapping from *beans* to *secrets* is still needed.[13] Note again that this is the wrong way around, as there seems to be an independent metaphorical connection between *spill* and *reveal* for some speakers, but not for *beans* and *secrets*, as was shown in Section 2.2.5.

In the case of idioms based on active metaphors it is also desirable to express the more general metaphors, in order to be able to deal with productive substitutions of words in these idioms. For example if the more general metaphor underlying the idiom *break the mold* is expressed as a higher-level constraint, then productive uses

[12] I consider 'word play' to be uses of idioms that involve reference to the literal meanings of the idiomatic words. A clear example of this is *let the neon cat out of the cellophane bag*. Coordination of idiomatic and non-idiomatic material such as *bite the bullet and the bread* is another example. Creative substitution of related lexical items like *panthers* for *wolves* might also fall under the definition, as it requires understanding the literal meaning and finding related lexical entries. But whether this type of substitution is acceptable depends very much on the type of word involved and its semantic relationship to the substituted word. In general it seems to be much more acceptable for the verbal part of an idiom than for the nominal part.

[13] One might think that the mapping is not truly 'independent' if this lexical entry for *beans* is only ever used with *spill*. However, this is a meta-level fact about the grammar as a whole, i.e. something that can be inferred, not something that is encoded as knowledge explicitly.

like *shatter the mold*, *crack the mold*, and *break out of the mold* can be understood based on the same metaphor.

2.2.8 Locus for Semantics of Non-Decomposable Idioms

Approaches to idioms also need to deal with the difference between semantically decomposable and non-decomposable idioms (Nunberg et al. 1994). This difference is clear in the corpus study, where decomposable idioms show much more variation than non-decomposable idioms (25% vs. 3%).

Typical non-decomposable idioms are *saw logs*, which roughly means 'snore', *shoot the breeze*, which roughly means 'chat', *make tracks* which roughly means 'leave', and *kick the bucket*, which roughly means 'die'. The reason that the meaning of *kick the bucket* is not decomposable is that 'die' is a one-place relation in which no complement like *bucket* plays a role. That is, there is no way to analyze the meaning of the idiom such that parts of it can be assigned to the individual words in the idiom. Because the meaning 'die' is not associated with any of the individual words, the phrasal pattern as a whole is needed, so it can carry this meaning.

One might think that the verb *kick* could carry the idiomatic meaning. However, this is not only unintuitive, but would also require changing how linking works, because otherwise the *i_kick_bucket_rel* would end up having the *bucket*'s empty meaning as its UNDERGOER, like all transitive verbs. Furthermore, this type of analysis would predict that the idiom can passivize.

2.2.9 More than Head-Argument Relationships

Idioms sometimes involve fixed items that go beyond mere head-argument relationships like those that are usually expressed via valence constraints.[14]

[14]This is also true for proverbs that are not completely fixed, like *the early bird catches the worm/early birds catch the worm/the early bird gets the worm...* or *that's the way the cookie crumbles/that's not the way the cookie normally crumbles/that's not just the way the political cookie crumbles*. A mechanism like the one developed in this dissertation could also be used for these proverbs, but I will not discuss whether or not that would be desirable.

Adjectives and Specifiers

Idioms can include adjectives and specifiers:

(32) a. bark up the **wrong** tree ('follow the wrong course')
b. vote a **straight** ticket ('cast a ballot uniformly for one party')
c. have an **itchy** palm ('ask for a tip')
d. roll out the **red** carpet ('provide special treatment')
e. get **one's just** desserts[15] ('get what one deserves')
f. give NP **some** skin ('touch hands in a special greeting')

These idioms are problematic for word-level syntactic and semantic approaches, because there is no reliable way to locate these adjectives and specifiers within the NP when variation and further modifications are possible.[16] They also contradict the claim in Baltin (1987) that idioms involve only the head of a phrase and the head of one of its complements.

Adverbs and Adjuncts

Idioms can also include adjuncts, such as adverbs and PPs that are not complements. Some examples are given in (33). Webelhuth (1994) gives similar types of idioms for German.

(33) a. (to) put it **mildly**[17]
b. be born **yesterday**
c. bend over **backwards** to do something

[15]The historically correct form of this idiom is *get one's just deserts*, but *get one's just desserts* is actually used about twice as frequently in the corpus. This is presumably because *desserts* is pronounced in the same way and is much more frequent. In fact, for many speakers *deserts* with this meaning is not an independent noun, and for these speakers even *get one's just deserts* is an idiom and not a collocation.

[16]Nouns do not usually specify what modifiers they can occur with, so they do not have a feature with which to express these constraints. This is discussed in more detail in Chapter 4.

[17]See Chapter 3 for a discussion of why this is an idiom.

d. divide/split something **fifty-fifty**[18]
e. skate **on thin ice**
f. cross a bridge **before NP comes to it**
g. look at something **through rose-colored glasses**[19]
h. build castles **in the air**
i. read **between the lines**
j. fall **on deaf ears**
k. hide one's light **under a bushel**

Information about adjuncts is not available in word-level approaches, unless one thinks all adverbs and other adjuncts that occur in idioms should be analyzed as complements or are otherwise available in the valence of verbs. These data are also incompatible with the claim in Coopmans and Everaert (1988) that idioms can only involve the direct θ-role of their head.

No Verb or Other Head

Some idioms do not involve a fixed verb. For example, the idiom *from/out of the frying pan into the fire* often occurs with the verb *jump*, but it is also found with other verbs like *be*, *go*, *leap*, *move*, *step*, *get*, *throw*, and *take*, or without any verb, as in (34). This idiom is also listed without a fixed verb in both idioms dictionaries.[20]

(34) a. He's out of the frying pan and into the fire. (The New York Times, 6/13/1996, p. A12)
b. [...] but North went from the frying pan into the fire. (The New York Times, 7/31/1997, p. C13)

[18]If the adverb *fifty-fifty* is used with other verbs, then this expression does not meet my definition of idiom. But this is not the case in the corpus, and at least for some speakers sentences like **we paid for the dinner fifty-fifty* are not acceptable.

[19]There appears to be some (dialectal?) variation concerning this idiom. 85% of the occurrences of this idiom in the corpus are of the form *rose-colored glasses*, but the rest are *rose-tinted glasses* and *rose-colored spectacles*.

[20]The Collins COBUILD dictionary lists it as *from the frying pan into the fire*, and NTC's American Idioms Dictionary lists it as *out of the frying pan into the fire*.

c. We are always advised not to leap from the frying pan into the fire [...] (The New York Times, 2/9/1995, p. C16)
d. Passengers who shift to foreign airlines, he said, may be moving "out of the frying pan into the fire." (The New York Times, 4/4/1989, p. D5)
e. [...] this is only stepping out of the frying pan into the fire. (The New York Times, 1/27/1985)
f. Sometimes you can get out of the frying pan into the fire. (The Tampa Tribune, 1/21/1998, p. 6)
g. [...] which would throw the American people from the frying pan into the fire. (The New York Times, 10/16/1981, p. A34)
h. [...] that might have taken the partnership from the frying-pan into the fire. (The New York Times, 11/18/1999, p. E6)
i. It may be a case of out of the frying pan and into the fire for William E. Yingling [...] (The New York Times, 8/1/1991)
j. "Out of the frying pan into the fire" is the motto being illustrated in this Dutch engraving (The New York Times, 2/2/2001, p. E33)

To take an example that is somewhat more frequent in the North American News Text Corpus (the examples in (34) are from Lexis-Nexis), out of the 41 occurrences of the idiom *butterflies in one's stomach* in the corpus, only 15 (37%) have one of the verbs *have*, *get* or *be* in them, and *have* and *get* are equally frequent, with 7 examples each.

(35) a. I've had butterflies in my stomach ever since I heard about this.
b. That's why, for instance, you get butterflies in your stomach when you get nervous.
c. There were butterflies in my stomach
(from the North American News Corpus)

Most of the other occurrences do not involve any verb at all, as can be seen in (36).

(36) a. [...] I would expect him to go out there with a little anxiety, a few butterflies in his stomach.

b. [...] men and women with hope in their hearts and butterflies in their stomachs sat silently [...]

c. Butterflies in the stomach are common the minute anyone has to sign on the dotted line [...]

d. Butterflies in my stomach, I moved into the current.
(from the North American News Corpus)

Except for (36d), where *butterflies in one's stomach* is a *with*-less absolute (see Chapter 6), it is a complete NP (or at least N′) in these examples.

One might think that the preposition *in* could be used to carry the representation for this idiom. This is quite unintuitive as it is clearly not the preposition *in* that makes this expression idiomatic, and it would result in the proliferation of lexical entries for *in* for each idiom of this type. Note also that it is not clear that the syntactic relationship between *in* and *butterflies* is the same in all these examples. In the NP examples *butterflies* is the MOD value of *in*, so the relationship could be expressed that way. But this relationship does not hold for the examples involving the verbs *have*, *get*, or *be*. If anything, the prepositional phrase would modify the whole VP in these examples, and *butterflies* is not the head of that VP. In some approaches, such as the one implemented in the LinGO grammar, there is even a difference in what the preposition's MOD and SUBJ values are for examples involving *have* vs. examples involving *there are*. In the latter case the value of MOD is not even a phrase that contains *butterflies*.

So, in order to make this type of analysis work technically, one would have to make the assumption that there are predicational prepositions that have a SPR feature which has *butterflies* as its value in examples where *in one's stomach* occurs with the copula, as the complement of other verbs, as a modifier of VPs, and as a modifier of nouns.[21] I am not sure whether there are any independent motivations for this.

Note also that there are some idioms of this type, such as *cat out of the bag*, where the set of verbs these can occur with is restricted. While both *let the cat out of the bag* and *the cat is out of the bag* are frequently used, it is not the case that just any

[21] It looks like the analysis in Sag and Wasow (1999) might be consistent with such an interpretation.

verb can be used: *?help the cat out of the bag*, *#throw the cat out of the bag*. But prepositions like *out of* presumably do not have any way of specifying which verbs they can occur with, so that such constraints cannot be expressed in a word-level approach lexicalizing the idiom in the preposition.

For similar reasons it is not possible to treat *butterflies* as the head of the idiom. While *butterflies in the stomach* is an NP in some examples, it is probably not a constituent in its canonical form and in many others because the verb (*have*, *be*, *get*, etc.) is missing from the smallest constituent that contains the whole idiom. This may be a little controversial as in some analyses *butterflies in the stomach* is a constituent even in these cases. However, uncontroversial examples involving locative verbs with obligatory prepositional complements can be found as well:

(37) a. [...] to do it in front of the president is going to put a few butterflies in my stomach. (The Houston Chronicle, 01/09/1998)
b. It puts a few butterflies in my stomach, a little edge (The San Francisco Chronicle, 12/31/1997)
c. [...] the City Hall chambers have put butterflies in many a stomach. (The Washington Post, 11/05/1997)
d. [...] the thoughts are putting butterflies in their stomachs. (Los Angeles Times, 06/29/1995)
e. It puts butterflies in your stomach to think of it. (The Seattle Times, 04/17/1991)

Furthermore, there are even examples where *butterflies* and stomach are not adjacent:

(38) a. Butterflies will be in Marcus Allen's stomach. (The Kansas City Star, 08/30/1998)
b. The butterflies will be in the stomach this morning (The Times, 06/16/2001)
c. [...] the butterflies are starting in my stomach. (Sunday Times, 01/28/2001)

Note that the idioms discussed in this section also contradict the following claim in Koopman and Sportiche (1991):

> If X is the minimal constituent containing all the idiomatic material, the head of X is part of the idiom

If this claim is meant to apply to every occurrence of an idiom there are many further counterexamples to this claim, such as examples involving raising.

Syntactic Constructions

Many syntactic theories, following Chomsky (1985), strive for a kind of modularity where statements of grammar (rules, constraints, or principles) refer only to general grammatical items (e.g., features or configurations) and the constructions discussed by traditional grammarians are considered epiphenomena. The data for idioms discussed in this dissertation is one type of evidence for an alternative conception of grammar in which constructions have primary ontological status. As Fillmore and Kay (1997) and Goldberg (1995) have shown, there are syntactic constructions that carry meaning which cannot be assigned to any of their parts. In such constructions, almost anything can be fixed.

(39) a. The bigger the better.
b. Pat sneezed the foam off the cappuccino.
c. Pat knitted her way across the Atlantic.

In principle any 'meaning contributed by the construction' could in many cases also be associated with a verb or class of verbs (derived by lexical rule) in the hierarchical lexicon, if such a verb is part of the construction. But that does not always seem to result in an intuitively sensible verb meaning.

The WXDY construction (Kay and Fillmore 1999), e.g. *What are your feet doing on the table?* is a good example: if it is to be given a word-level representation, it has to be represented as a special lexical entry for *be*. That is both unintuitive and problematic: the 'incongruous' aspect of the meaning would have to be located on *be*, which would result in the question meaning taking scope over this aspect of the meaning. This is discussed more fully in Chapter 5.

It is true for many other constructions as well that their meanings are not easily associated with their parts, e.g. *X is nothing if not Y* ('X is really Y', 'X is certainly Y'). (40) does not mean that Hugh Grant is nothing if he is not charming, and no part of this expression directly carries the 'really', 'certainly' aspect of the meaning of this construction.[22]

(40) Hugh Grant **is nothing if not** charming.
(from the North American News Corpus)

Another construction was studied extensively by Zwicky (1982), who found that 57% of his 157 Verb Phrase Deletion cases with stranded infinitival *to* involve the verbs *want*, *have*, and *used*, and 82% involve subject-controlled predicates,[23] while many other types of infinitivals never occur in this construction. This is confirmed in the NYT part of the corpus where there are no occurrences of *hard to* or *whether to* stranded before a period or comma, while there are about 15000 occurrences of non-stranded *hard to* and about 8000 occurrences of non-stranded *whether to*. Some typical corpus examples of this construction are:

(41) a. I can give up smoking whenever I want to.
b. We'll deal with that when we have to.
c. Pete didn't play the way he used to.
d. If we want to work in Ireland we will be able to.
(from the North American News Corpus)

The frequency of *want to*, *have to* and *used to* cannot be explained by the overall frequency of these verbs in infinitival constructions, either. I studied *used to* and *able to* in more detail and found that *used to* is about[24] 12 times more likely to get

[22]See Chapter 3 for a discussion of why the literal meaning plus a Gricean rule of interpretation is not sufficient.

[23]A quick corpus search for *to* followed by a comma or period in the NYT part of the corpus roughly confirmed those findings, the percentages being 52% and 85% respectively. As it produced over 11000 matches, the elimination of occurrences of the preposition *to* was done only on a word-by-word basis based on the preceding word, e.g. eliminating all occurrences of *talk to* and *walk to*, and parentheticals were not eliminated from the data.

[24]The actual difference may be a bit different as not all occurrences of the preposition *to* were eliminated carefully.

stranded before a comma or period than *able to* is (2.80% vs. 0.24%), and 12 times more likely to be stranded before a period (1.59% vs. 0.13%). In other words, while *able to* is about twice as frequent in the corpus as *used to*, there are about 6 times more stranded occurrences of *used to* than of *able to*. This can be seen in Table 2.2.

	stranded before a period	stranded before a period or comma	total # of tokens
used to	311	546	19524
not to	227	318	37695
ought to	34	54	5184
able to	50	97	39818
hard to	0	0	14869
whether to	0	0	7901

Table 2.2: Stranded and Non-Stranded Occurrences in the NYT Corpus

Zwicky (1982) also got judgment data on 21 key examples from 74 speakers, which shows that they have generalized from the core cases *want*, *have*, and *used* in different ways, resulting in different constraints on what they find acceptable. These three verbs do not form a natural class, although they have various properties in common, such as cliticizing (*wanna*, *hafta*, *useta*) and being modals/quasimodals. But there are other cliticized verbs and quasimodals (*oughta*, *gotta*, *gonna*). As can be seen in Table 2.2, *ought to* occurs stranded before a period 5 times more frequently than *able to* (0.66% vs. 0.13%), and 4 times more frequently before a period or comma (1.04% vs. 0.24%).[25] So the analysis of this construction is underdetermined by the data, and speakers differ in how they generalize. Some speakers are very conservative and stick closely to the properties of the most frequently encountered cases. Other speakers are more willing to generalize, but do so in different ways. For example, many speakers have a "VP constraint", i.e. they require the constituent immediately preceding stranded *to* (on which it leans phonologically) to be a VP or a predicator in a VP (i.e. V, *be* + Adj, or a verbal idiom). These speakers do not accept examples like *persuaded Clinton to* or *I wonder whether to*. Other speakers are more restrictive and

[25]It would be interesting to see how frequently *got to* and *going to* occur stranded—Zwicky's data suggest that they may occur stranded more frequently than expected as well. But this would require a careful study as there are likely to be many 'false hits'.

reject a preceding predicator when it is impersonal, as in *I think it's hard to*. Almost all speakers make exceptions for monosyllabic non-lexical items (like *not*, *how*, *him*). This is probably due to the fact that certain instantiations such as *not to* are more frequent than expected and are probably fixed units, and speakers have generalized from the properties of these frequent sequences (Zwicky 1982:45).

Zwicky's data shows that there are phonological constraints on this construction, and that the construction has lexicalized subsorts. It also suggests the construction is learned 'bottom up' from its more frequent incarnations. Note that Sag (1999) and Bender (2001) treat ellipsis as a construction as well. The existence of this kind of construction provides another piece of evidence for the need for phrasal construction types.

The information about whether something is a question, statement, or imperative clause may be seen as a more general type of construction meaning (Sag 1997). And as was discussed in Chapter 1, many general syntactic and semantic constraints are already associated with phrasal construction types. This system of hierarchical classification of phrases, which is already in place, can be used for idioms as well.

2.2.10 Interaction of Idioms and Syntactic Constructions

Data about the interaction of some idioms and constructions also point toward a need for a constructional approach. This is discussed in Chapter 6, where the *with* and *with*-less absolute constructions and their interaction with predicative idioms are studied.

2.2.11 Collocations

I use the term collocation for expressions made up out of two or more words that have one of their literal meanings, but are conventionalized in this combination. That means that they occur more frequently in this combination than would be expected from the frequency of their parts, compared to alternative expressions that could be used (Berry-Rogghe (1973), Firth (1957)). A typical example from the literature is *vanishing cream* vs. *disappearing ink*. A VP example is *bear the brunt of*—there are

more occurrences of this collocation (704) than of *brunt* outside of this collocation (481), making it a clear case of an unexpectedly frequent combination of words, especially compared to other combinations like *take the brunt of*, *suffer the brunt of*, *receive the brunt of*, etc., which occur as well, but not nearly as frequently, and *bear the force of*, *bear the majority of*,[26] and *bear the greater part of*, which do not occur in the corpus at all.

It is shown in Chapter 3 that some collocations, like idioms, have canonical forms. This also necessitates an approach that includes phrasal patterns. But one needs more than just a representation for the canonical form of these collocations. As with idioms, a more general representation is needed to capture the fact that the variations are related to the collocation, and to predict that they are more frequent and more likely contain the words in the same senses as in the 'canonical form' than would be expected by chance. For example 'hold (onto) one's turf' is not likely to mean 'grasp a piece of soil' even when varied, as in the corpus examples in (42).

(42) a. [...] he usually must pay a portion of the profits to organized gangs for the privilege of **holding his curbside turf**.
b. [...] trade groups in the Freedom to Advertise Coalition have been fighting in court to **hold onto their tobacco turf** [...]
c. Meanwhile, the big construction companies are trying to grow by making inroads into **turf traditionally held** by medium-size builders.
(from the North American News Corpus)

If a general representation for *hold ... turf* is lexicalized then the general principle that longer lookup is preferred will predict this. Note that *hold (onto) one's turf* is not an idiom according to my definition because *turf* has the same meaning (roughly 'territory') in *on one's own turf*, *this is my turf*, or *turf wars*.

[26]There is one example of *bear the majority of the body's weight*, which is presumably a different meaning of *bear*.

2.2.12 Psycholinguistic Evidence

This section discusses psycholinguistic evidence for phrasal patterns and other relevant findings which point towards a constructional approach. Most of them individually are consistent with a variety of approaches, but taken together they lend some support to the approach proposed in this dissertation. Note that I do not claim to have a complete psycholinguistically plausible model of HPSG, but only argue at a very general level what architecture seems to comport best with the available psycholinguistic evidence.

Idioms are processed faster than non-idioms

This has been found by several studies (e.g. McGlone et al. (1994)) that compared the understanding of idioms and literal phrases with comparable meanings, e.g. *spill the beans* vs. *reveal the secrets*. Familiar idioms were understood more quickly than their literal paraphrases. This finding is inconsistent with any approach (e.g. Pulman's) that requires the literal meaning of an expression to be constructed first. In the approach proposed in this dissertation, idioms would be expected to be understood faster since they are represented as (partially specified) units, requiring less computation than comparable phrases that have to be built from their parts. While this effect might not show up in all computational implementations of the approach, it is generally accepted that for humans, direct access from memory is very fast, and therefore it is faster to retrieve a complete phrase from memory than to retrieve its separate parts and to assemble them. This is also the explanation given by Everaert et al. (1995):

> The fact that idiomatic expressions are processed faster in their idiomatic sense than in their nonidiomatic sense then could be due to the fact that the meaning of idioms does not have to be computed but may be found in the lexicon, thereby reducing processing time. (Everaert et al. 1995:10)

Idioms in canonical form are understood faster than their variants

This was shown by McGlone et al. (1994), who found that idioms occurring in their canonical forms were understood more quickly than variants of these idioms. It has

a very natural explanation in an approach where canonical forms have their own representation as a subtype of the idiom, again speeding up processing because they can be retrieved as pre-assembled wholes.

Decomposable idioms are processed more slowly than non-decomposable ones

This has been shown by Gibbs and Gonzales (1985), who first established 'frozenness' by asking subjects to judge whether variations of idioms maintain their idiomatic meaning, and then found in a separate experiment that subjects were faster at processing 'frozen' idioms compared to flexible ones. This finding could be due to the fact that non-decomposable idioms do not require as much computation as decomposable idioms. This is because more information is present in the phrasal representations for non-decomposable items. Because they are less syntactically flexible, they can be more fully specified, and therefore require less computation.

Frequency and familiarity affect the understanding of idioms

Cronk et al. (1993) show that frequency and familiarity have large effects on the processing of idioms. Idiomatic sentences biased toward a figurative interpretation were presented word-by-word. The mean reading time per word in the idiom phrase was significantly faster when the idiom was high-frequency and high-familiarity. The familiarity levels of these idioms had been established in a separate experiment. This finding is not consistent with the approach in Pulman (1993) because there is no idiom representation to associate such information with. While it might be argued that frequency and familiarity are not grammatical properties of words and phrases, they do affect processing, and presumably correspond to varying degrees of entrenchment of the representations for these phrases in the memories of humans. Also, only if a representation of an idiom exists can it be annotated with probabilities in a computational system that uses such information to increase efficiency or model human performance.

Conventionality affects intuitions about transparency

Keysar and Bly (1995) show that after teaching people various meanings for unfamiliar idioms by presenting these idioms in contexts that bias different meanings and asking them to select the most plausible meaning from a list, they will perceive the learned meaning as more transparent and its opposite as less transparent, regardless of which meaning they were taught. For example, subjects who were taught that *to find an elephant in the moon* means 'to point out something that should have been obvious to all' rated that meaning as more transparent, and also tended to believe that that is also the meaning an uninformed person would be more likely to assign to that expression than 'to make a spurious discovery; an illusion'. Furthermore, in a second experiment in which some subjects were asked to use the learned idioms in two example sentences, these subjects rated the opposite meaning as less transparent than subjects that were not asked to use the idioms.

This finding is inconsistent with the position, taken by Gibbs in various papers, that the meaning an idiom is understood to have is determined by the existence of the underlying metaphor:

> We understand *let off steam* to mean 'release tension from anger' because there are underlying metaphors, such as THE MIND IS A CONTAINER, that structure our experience. (Gibbs and Nayak 1989)

It is not the case that this idiom is understood to have that meaning 'because' of that metaphor—the Keysar & Bly study shows that many other metaphors may have made just as much sense. The metaphor did probably motivate the idiom originally, and it is possible to see the metaphorical basis of an idiom after comprehending it, but the main reason we understand an idiom to mean what it means is because we learned its conventional meaning. If we had grown up with a language in which *let off steam* means 'produce output' we might think that this makes great sense, because a steam iron produces output when it lets off steam.[27]

[27]One might think that it is not plausible for an idiom to be based on steam irons, and it is true that in our current society steam engines play a far more important role, and are therefore more likely to give rise to an idiom. However, the effects of conventionality are such that *if* an idiom

2.3 The Problem of Variation

Almost all idioms are variable to some extent and cannot be seen as a simple fixed string of words. Nevertheless, it is desirable to list an idiom only once, and use independently existing mechanisms of the grammar to derive the variations, because they are clearly instances of the same idiom with the same meaning. Some types of variation, such as semantically internal modification, also show that the parts of some idioms have to carry parts of the idiomatic meaning.

2.3.1 Variants Differing in Inflection

Inflectional information for idiomatic usages of words is identical to that of nonidiomatic usages, and should not have to be repeated. Therefore any approach has to establish some sort of link to the non-idiomatic lexical entries. Almost all idioms can occur in a variety of tenses:

(43) a. Royal Housekeeper Spills The Beans
b. He had already spilled the beans in an intimate book.
c. Will Colombia's drug lords spill the beans?
(from the North American News Corpus)

Sometimes the nouns involved can be either singular or plural:

(44) a. There is an inherent fear that inflation is going to **rear its head**
b. Both problems have **reared their heads** again
(from the North American News Corpus)

And whenever this happens, the same regular or irregular forms are used that the nonidiomatic uses of the words have, as the examples in (45) show.

(45) a. He's a sheep in wolf's clothing.

has a certain meaning, speakers are apparently quite happy to find a justification for it even if an alternative meaning would make more sense to someone who has not been taught a conventional meaning for the same expression.

b. They are sheep in wolves' clothing.
c. *They are sheep in wolfs' clothing.

More generally, when words used in their literal meanings have irregular inflections it is not a coincidence that idioms involving these words use those same inflections:

(46) a. He got his feet wet./*He got his foots wet.
b. She held her tongue./*She holded her tongue.

This shows that the intuitively appealing idea that idiomatic phrases are simply learned as 'fixed expressions' and stored as unanalyzed strings must be wrong. Inflectional information for idiomatic usages of words should not have to be stated separately. A mechanism is required for relating the morphology of idiomatic and non-idiomatic usages of words, while making sure their semantics is not simply equated.

2.3.2 Open Slots

Some idioms have open slots into which, e.g., any *NP* can be inserted. Examples of this are *sell NP short*, *give NP some skin*, *stab NP in the back*, *send NP packing*, *nip NP in the bud*, *sweep NP under the carpet*, *take NP by storm*, *keep NP under wraps*, etc. These idioms are a problem for a phrasal phonological approach. As the examples in (47) show, some of these idioms do not have to occur as a fixed VP, so they are a problem also for phrasal syntactic approaches.

(47) a. And Ford was doing its best to **sweep under the carpet the almost unbelievable fact that it is abandoning its car manufacturing activities in the UK** [...] (The Daily Telegraph, 3/3/2001, p. 2)
b. The experience is that if you **sweep under the carpet these kinds of very serious crimes**, it is a very negative message for the future. (The New York Times, 12/31/1998, p. A1)
c. Quite simply, the Government's overriding objective is to win the election, and **the euro is an awkward topic that it would like to sweep under the carpet** until that objective is secured. (The Independent, 9/10/2000)

d. [...] **the issue will be hard to sweep under the carpet** [...] (Sunday Times, 1/10/1993)

e. As long as it's undefined, **it's easy to sweep under the carpet** and ignore. (The Toronto Star, 5/23/1991, p. A6)

In other idioms the specifier can be any possessive pronoun or NP: *whet someone's appetite*, *call someone's bluff*, *force someone's hand*, *seal someone's fate*, *cross someone's path*, *breathe down someone's neck*, *do someone's dirty work*, etc. Some of these idioms allow *of* complements instead, as in (48):

(48) In announcing that they will roll out two online music subscription services this summer, the five biggest record labels have **whet the appetites of consumers and investors**. (The New York Times, 4/23/2001, p. C1)

In some idioms like *blow one's cool*, *bury one's head in the sand*, and *burn one's bridges*, the possessive pronoun has to agree with the subject of the idiom.

Open slots are not limited to these particular positions, as examples like *play cat and mouse with NP*, *fly in the face of NP*, and *NP's days are numbered* show. All these idioms are a problem for phrasal and word-level phonological approaches, and they contradict Gestel's (1992) claim that idiomatic phrases contain all and only fixed material.

2.3.3 Modification

As Nunberg et al. (1994) note, the parts of some idioms need to be modifiable in the syntax, even for non-decomposable, fixed idioms. One type of modification is the meta-linguistic type using adjectives like *proverbial*:

(49) Even the stoic Encyclopedia Britannica **hit the proverbial ceiling**

But some other adjectives can be used with non-decomposable idioms as well:

(50) a. We're initially drawn into them by the discovery of corpses and the question of who made them **kick their respective buckets**.

b. A friend of mine whose husband has been **sawing some major logs** every night of their marriage has tried everything.

c. Fields was furious with Landrieu for not **closing Democratic ranks** and endorsing him in that runoff.

d. In the Senate Chris Dodd, Bill Bradley, John Kerry, John McCain and Al D'Amato love to **shoot the on-air breeze** with the craggy-faced shock jock.
(from the North American News Corpus)

The examples in (49) and (50) can be thought of as external modification in the sense that the scope of the adjective is external to the idiomatic VP. That is, the modifiers modify parts of the idiom syntactically but not semantically, and are therefore compatible with non-decomposable idioms in which the nouns they modify do not carry part of the meaning of the idiom. For example, the relevant parts of (50a-b) mean roughly 'who made them die, respectively' and 'snoring in a major way'. A mechanism for achieving such wide scope is independently needed for examples like *an occasional sailor walked in* or *drink a quick cup of coffee*.

Nevertheless, one may insert these adjectives syntactically, which is not possible in any approach which fixes the phonology or fixes the syntax phrasally. Even word-level approaches which locate things via the constituent structure (i.e. the DTRS) have a problem, since the head noun is in a different place in the syntactic structure when it is modified—see Section 4.1.3.

Some specifiers, adjectives, and compound parts can even internally modify parts of idioms (Ernst 1981). This is possible only with decomposable idioms, in which the individual words carry parts of the meaning. I found corpus examples of this type for all the semantically decomposable idioms I studied—see Chapter 3. Some examples with adjectival modification are given in (51):

(51) a. King and Alexander, who sued each other after their bitter 1992 divorce, **buried the legal hatchet** in May.

b. But Herzog pleaded he had been misunderstood and critics had a hard time tagging him as "far-right" because of his other views that **broke the usual**

mold of nationalist thought.

c. After weeks of vacillating, many socialists seem to hope that Papandreou himself will **call the final shots**.

d. Meanwhile modern navigation and transport ensured that **no significant stone on the planet was left unturned**, no nation or tribe undiscovered or undocumented.
(from the North American News Corpus)

The relevant parts of these examples do not mean 'legally reconcile the disagreement' or 'in the legal domain, reconcile the disagreement'; 'usually escape the pattern' or 'in the usual domain, escape the pattern'; 'finally make the decisions' or 'in the final domain, make the decisions'; 'significantly, check all places' or 'in the significant domain, check all places'. Instead they mean 'reconcile the legal disagreements, 'escape the usual pattern', 'make the final decisions', and 'check all significant places'.

Note that for many examples it is not straightforward to decide whether they are cases of external or internal modification, because both interpretations are plausible. For example, *turn the political tables* can have an internal interpretation 'reverse the political positions', but also an external interpretation 'in the domain of politics, reverse the positions'. It is not clear whether there is any difference between these interpretations, or how to decide which interpretation the writer intended. There is controversy on both sides of this issue. Some researchers, like Nicolas (1995), claim that semantically internal modification does not exist at all, while other linguists argue that there are at least some clear examples of internal modification, and want to treat as many cases as possible as internal modification, considering that these readings are more natural, and a precise formal approach for handling external modification has yet to be developed.

There are also many examples of unusual specifiers used with decomposable idioms. Note that it would make no sense to quantify over something that is not there, so examples involving quantifiers show that the idiomatic nouns must have meaning associated with them.

(52) a. We are trying to **break that mold** now.

b. His U.S. experience and fluent English should help **break some of the ice** with President Bill Clinton [...]
c. Canon's new Elan II-E (list $800) is **making a lot of waves** in the photo press [...]
d. [...] corporate sponsors still **call most of the shots**.
e. [...] everyone **loses too much face** by backing down [...]
f. [...] Russia cannot be allowed to **call NATO's shots**.
g. [...] it seemed reasonable for Mr. Clinton's advisers to assume the study would **sound the B-2's death knell**.
(from the North American News Corpus)

There are also quite a few examples of idiomatic nouns being used as part of compounds:

(53) a. [...] members once could **call the public policy shots** at the ballot box
b. [...] the firm hoped to **make marketing waves** with Ms. Hirsch's celebrity status and stock-picking prowess.
c. The president and pro-choice activists were able to rally support from pro-choice voters, with rhetoric suggesting that extremists were **running the GOP show**.
d. South African President Nelson Mandela's wardrobe of colorful shirts has made him perhaps the first statesman to **break the suit-and-tie mold** during his official visit to Britain.
e. Russia and NATO, **breaking post-Cold War ice**, got down to details Sunday
(from the North American News Corpus)

Idiomatic nouns can also be modified by relative clauses and *of*-phrases:

(54) a. Jews, Communists, Catholics, Freemasons, Mormons, international bankers, the CIA, the Trilateral Commission, the Council on Foreign Relations - all have been accused of plotting takeovers or **pulling strings that control political or economic decision-making**.

b. The Robinsons have **buried the family hatchet that made the band's 1994 album, "Amorica," such an angry and depressing work**
c. Dr. Hall has called middle schools a weak link in the educational chain where too many children **lose ground that they will be unable to make up in high school**.
d. [...] the pope showed he knows how to **strike a chord that reaches their emotions**.
e. [...] the decree has attempted to **level the playing field that IBM once dominated** [...]
f. [...] the practical result will be to **level the playing field that currently favors big business**
g. I hope that this day will **break the ice of mistrust and hostility** which exists in both Latvia and Russia
h. A Chirac-Balladur runoff would **break the mold of recent presidential elections here** which have wound up as duels between a conservative and a Socialist.
(from the North American News Corpus)

Idioms containing more than one idiomatic noun provide even clearer examples. Consider the examples in (55):

(55) a. Next year will be the year **the information wheat is separated from the digital chaff**
b. But the Senate Finance Committee **let the strategic cat out of the fiscal bag** when it released a new study
c. With **so many financial skeletons** having tumbled out of **the corporate closet**, it may take some time for skittish investors to regain confidence in the company.
d. Palestinian President Arafat, **caught between the rock of Israeli policy and the hard place of a population in deepening need**, had pressed Israel to relax the closure ahead of the summit.

e. They're meant to make it possible to find **that needle of information in the electronic haystack of pages out there**.
(from the North American News Corpus)

In examples like these, where two nouns are modified, it is not possible for both of the modifiers to get a domain modification analysis.

2.3.4 Passive

The pieces of many idioms can appear separated from each other in passive constructions.[28]

(56) a. **The moral tables** have been **turned** [...]
b. [...] **the shots** clearly will be **called** from San Francisco [...]
c. [...] her activism in a slot where few **waves** have been **made** since the late 1970s.
d. **The death knell** has been **sounded** many times for newspapers.
e. If **the mold** could be **broken**, it would be helpful for everybody [...]
f. **The real emotional chord**, though, was **struck** with Yad Vashem [...]
g. But **the ice** was later **broken** at the following plenary session [...]
h. **The playing field** has been **leveled** through redistricting [...]
(from the North American News Corpus)

This is not a problem for any word-level approach because the passive lexical rule could apply normally. But it rules out phrasal approaches which fix the phonology or syntax of the phrase, unless they allow subsequent transformations. In the past, this has often led to the assumption that idioms must be represented at the word level or at some pre-transformational level. But passivization is not a problem for phrasal approaches like the one proposed in this dissertation, which specify only the semantic relationships between the parts of the idiom and not their syntactic configuration (see Chapter 5).

[28] I also found some examples of this in the German part of the ECI/DCI Multilingual Corpus I. 3 out of the 26 occurrences of the German idiom *die Katze aus dem Sack lassen* ('let the cat out of the bag') were passive.

2.3.5 Topicalization

I did not find any examples of topicalization in the corpus for the idioms I studied,[29] so the examples below are from Nunberg et al. (1994).

(57) a. Those strings, he wouldn't pull for you.
b. His closets, you might find skeletons in.
c. Those windmills, not even he would tilt at.
d. That hard a bargain, only a fool would drive.

2.3.6 Distribution Over Several Clauses

As McCawley (1981) observed, parts of idioms can be spread over a main clause and a subordinate clause:[30]

(58) a. [. . .] **all the strings that Sounders coach Alan Hinton pulled** to get Baggio to make an appearance [. . .]
b. But the governor's biggest contribution may be **the close tabs he keeps on his ground operation**.
c. [. . .] the United States is trying to regain **face lost** when President Clinton backed down last year [. . .]
d. **The waves Japanese authorities are making** in the currency markets could see the Swiss franc supplant the yen as a favored source of cheap funding

[29]Presumably topicalization in general is fairly rare in the newspaper corpus so that the absence of such examples in the data I studied is not significant.

[30]These data are actually part of a paradox for transformational approaches to idioms discovered by McCawley (1981). The paradox arises under the assumption that idiomatic elements have to be adjacent in D-structure. It might be possible to explain (58a), if it is assumed that *the strings* are adjacent to *pull* at D-structure and raised out of the relative clause. But then sentences like *Pat pulled the strings that got Chris the job* should not have an idiomatic interpretation, since *the strings* would not be in the main clause at D-structure. There does not seem to be one single set of assumptions about movement that can accommodate the acceptability of both these examples. However, for non-transformational approaches it is clear that only examples like those in (58), i.e. those that are spread over a main clause and a relative clause, are potentially problematic.

e. The singers and their pastor have no doubts about **the dividends their presence has paid** for the families of the victims.
f. [...] defy **the death knell much of Wall Street was sounding** a year ago.
g. [...] **a mold that badly needs breaking**.
h. [...] **the popular chord that Buchanan has struck** [...]
i. He mourns **the backseat that books have taken** to movies [...]
(from the North American News Corpus)

This is a problem even for a word-level semantic approach, since the relative pronoun[31] does not meet the subcategorization requirement—the INDEX is shared between it and the modified noun, but not the semantic *relation*. In Sag (1997) even in bare relative clauses only the INDEX is shared. This is discussed in more detail in Chapter 5. The only approaches that can handle these data are semantic underspecification at the phrasal level, the TAG approach, and Pulman's inference approach. This is discussed in detail in Chapter 4.

2.3.7 Other Variations

Some idioms can participate in raising constructions:

(59) a. The hatchet now **appears** to have been buried for good [...]
b. [...] there are a lot of loose ends we **need** to tie up [...]
c. [...] there aren't any loose ends that **need** to be tied up [...]
d. [...] the tables **appeared** to have turned.
e. Wednesday, the tables **appeared** to turn.
f. But Clinton's is a victory where the tables **seem** utterly turned from 1976.
g. But Baucus, who supported the reintroduction of wolves, **wants** better tabs kept on the wolves.

[31] Sag (1997) argues convincingly that *that* should be considered a relative pronoun. Note that these examples are also acceptable with indisputable relative pronouns like *which*.

h. Sooner or later the piper **must** be paid.
(from the North American News Corpus)

For idioms like *pay the piper* in which the complement refers to an animate entity, control (equi) constructions are also possible (Gazdar et al. 1985:241).

(60) a. To some, the lesson is that no matter how happy the tune may be, eventually, the piper wants to be paid. (St. Petersburg Times, 3/2/2000, p. C1)
b. "The piper wants to be paid, needs to be paid, for the valuable service he renders to us all." (The New York Times, 3/19/1987, p. C21)

Some idioms can also occur in inchoative alternations:

(61) a. But the tables have now turned.
b. The ice has broken.
c. [. . .] the death knell sounded for the diesel-powered American car at the end of the 1985 model year.
(from the North American News Corpus)

Other types of variations, such as clefting, pied-piping, free relatives, and existential *there* constructions are also possible with some idioms:

(62) a. [. . .] he finishes off by strapping himself to a harness and pulling the truck a half-mile. Sure that's not the readers' leg you're pulling, George?
b. [. . .] guerrilla and other groups on which the United States wants to keep tabs
c. [. . .] we really haven't seen the goods delivered [. . .]
d. [. . .] which nonetheless paid what Brown and her advisers believe are important dividends.
e. Bossi sounded what seemed to be the death knell for Berlusconi's seven-month-old government
f. there is a piper to be paid some time

g. There was certainly ice to be broken
(from the North American News Corpus)

Some idioms occur in other variations:

(63) a. put the cat among the pigeons
b. the cat is among the pigeons
c. let the cat out of the bag
d. the cat is out of the bag

Similar examples were discussed in Section 2.2.9, focusing on the problems these idioms cause for word-level approaches because there is no suitable place at the word level where the relevant relationship could be stated. But this is also a type of variation, and as such provides another piece of evidence that idioms cannot be represented as complete, fixed constituents.

2.3.8 Pronominal Reference

Idiomatic nouns that are part of decomposable idioms can serve as antecedents to pronouns as in (64).

(64) a. No soap opera worth its bubbles would spill all the beans in one episode if it could dribble **them** out over many.
b. But a determined campaign by Garcia [...] and the entry of Cantu Rosas two weeks after the filing deadline in late September, leveled the playing field and may now have tipped **it** to Garcia
(from the North American News Corpus)

VP ellipsis as in (65) is also possible, although I have no corpus example of this type:

(65) I thought the beans would be spilled, but they weren't.

Pronominal reference to parts of idioms is further evidence for the fact that these parts are associated with parts of the meaning. Note that this causes additional problems for those approaches that try to prevent idiomatic words from occurring outside the idiom by making them have no meaning.

2.3.9 Incomplete Idioms

Some idioms can be recognized when only some of their parts are present. For example, the idioms in (66) can be recognized in the examples in (67), taken from Pulman (1993), even though they are incomplete.

(66) a. count one's chickens before they are hatched
b. scrape the bottom of the barrel

(67) a. That's a case of counting your chickens.
b. That suggestion came from the bottom of the barrel.

No approach that I am aware of is able to handle these data, and at the same time prevent random parts of idioms from occurring in literal sentences as in (68).

(68) #He revealed some very interesting beans to me yesterday.

It is formally possible to allow for these particular examples by making entries for the parts of these idioms which can occur by themselves. This may be justified for some idioms like *bottom of the barrel*.[32] But such a stipulation does not explain in general why a particular idiom allows this. Frequency information may help somewhat, because the frequent occurrence of idiom parts like *counting one's chickens* is expected if they have become separate idioms. Probably the amount of information present in the part that can stand by itself also plays some role. It can be expected that this phenomenon is facilitated for parts of idioms that are easily recognizable. For example, it is unlikely that an idiomatic word like *tables*, which is a fairly high-frequency lexical item in its literal meaning, would be sufficient for anyone to recognize the idiom *turn the tables*, while *death knell* is infrequent enough in its literal meaning for speakers to recognize its relationship to *sound the death knell*. It is also more likely that a larger chunk such as *counting one's chickens* is used by

[32]A corpus search found that *bottom of the barrel* actually occurs much more frequently without than with *scrape*. There are 57 such occurrences, 47 of which are in their canonical form *bottom of the barrel*. There are 18 examples of *scrape the bottom of the barrel*, 15 of which are in their canonical form.

itself because it is easily recognizable. But it does not seem possible to predict the (im)possibility of the creative use of novel incomplete idioms in a formally precise way. For example, as mentioned above, many idiomatic words such as *beans* cannot be used in this way (#*I have some interesting beans to tell you*).

These examples are like allusions and probably require some active reasoning on the part of the listener. It might be possible to build a system that is pretty good at 'interpreting' these allusions, by noticing the similarity to some fixed expression and replacing the relevant parts in that expression's meaning. But making sure that a generator produces only natural-sounding allusions would require large amounts of world knowledge and at least some level of 'creativity'. However, a theory of allusions is needed not only for idioms but also for quotes, proverbs, etc., so this is not an issue that needs to be addressed within the theory of idioms. A theory of allusion should be able to deal with incomplete idioms in a way analogous to incomplete quotes.

Note that there are some cases like *death knell* and *count one's chickens* which may be productive allusions to complete idioms for some speakers while other speakers have independent lexical entries for them.

2.4 Summary

In this chapter I have shown that there are many reasons to represent idioms at the phrasal level, but that there are also many types of variability that make it impossible to represent them as syntactically fixed phrases. Together these two types of data show that an approach that represents idioms at the phrasal level and expresses the relationship between the parts of the idiom semantically is the only approach that meets all the requirements of the data. Note that such an approach is exactly what Nunberg et al. (1994) call for:

> A central feature of our analysis of idiomatic combinations is the claim that the dependency between the verbs and their objects is semantic in nature, that is, that the inability of idiomatic *the beans* to appear with any verb other than *spill* is derived from the fact that the idiom

> consists in a (literal) 'spilling-the-beans' meaning being conventionally and homomorphically associated with a 'divulging-the-secret' meaning. (Nunberg et al. 1994:505)

Chapter 3

A Corpus Study of Canonical Forms and Variation

Why a corpus study? Some people might think that examples of idiom variation, even those found in corpora, are infrequent and can be set aside as "word play". Others might think that because some examples of variation are acceptable, idioms have to represented at the word level, and are just as likely to occur varied as combinations of other lexical items are. The corpus data in this chapter show that—in contrast to nonidiomatic combinations of words—idioms have a strongly preferred canonical form, but at the same time the occurrence of idiom variation is too common to be ignored.

I study four sets of data: Section 3.2 studies some idioms that have been discussed in the literature, Section 3.3 studies some idioms that are relevant for my argumentation, i.e. those that involve something other than the relationship between a verb and the head of its nominal complement, Section 3.4 studies some idioms involving words that have no literal counterpart, and Section 3.5 studies a randomly selected sample of V+NP idioms. I also compare the results to a few comparable non-idioms as a 'baseline'.

The corpus data reflect the difference between semantically decomposable and non-decomposable idioms (Nunberg et al. 1994)—the latter exhibit a lot less variation, both in terms of percentages of varied examples and in terms of types of variation

observed.

3.1 Methodology

The data in this chapter come from a large newspaper corpus of English, the North American News Text Corpus from the Linguistic Data Consortium (LDC).[1] This corpus contains approximately 350 million words. It consists of 5 sub-corpora: 173 million words from the New York Times News Syndicate (7/94–12/96), 40 million words from the Wall Street Journal (7/94–12/96), 52 million words from the Los Angeles Times & Washington Post (5/94–8/97), and 85 million words from Reuters News Service (4/94–12/96), separated into Reuters General News and Reuters Financial News.

Many sentences occur in this corpus repeatedly. One of the reasons for this is that press releases from Reuters are frequently re-released in slightly modified form, and they and New York Times syndicated articles are often used in more than one newspaper as well. Therefore I eliminated all duplicate sentences automatically before classifying and counting the examples. While this clearly improved the overall reliability of the results, it is possible that some short sentences were actually used twice and should not have been deleted. There was no way to identify these. This method also did not catch slightly modified duplicates, although I eliminated many of them manually when I noticed them because they were adjacent to each other in the sorted files.

The initial set of data for each idiom was obtained by using the Stuttgart Corpus Workbench.[2] For a typical V+NP idiom like *spill the beans* this involved searching for all sentences which contain both a form of the verb *spill* (expressed as a disjunction) and a form of the noun *bean* (also expressed as a disjunction). This was repeated for the 5 sub-corpora. Then the results were concatenated and the line numbers deleted (with an emacs keyboard macro) to make identical sentences identifiable as identical lines. Then the UNIX commands 'sort' and 'uniq' were used to eliminate duplicate

[1] http://www.ldc.upenn.edu/

[2] http://www.ims.uni-stuttgart.de/projekte/CorpusWorkbench/

lines. A preliminary classification into canonical and non-canonical data was made using 'grep'.

The queries looking for the two words anywhere in the same sentence, in either order, sometimes produced large numbers of irrelevant matches, depending on the frequency of the individual words involved. The next step involved manually deleting all these 'false hits' where the noun is not even a complement of the verb. For example when studying the idiom *call the shots*, an irrelevant match for *call* and *shots* in the same sentence would be *She heard shots and called 911.* The final classification into literal and idiomatic uses and into the various types of variation data was also done manually, although 'grep' and 'wc' were sometimes used to do further pre-sorting and to check the count, especially with large sets of data.

Note that for each idiom I studied only one of the conventional variants—the one listed as the main one in NTC's American Idioms Dictionary and/or the Collins COBUILD Dictionary of Idioms.[3] For example I studied only *hit home* and not *strike home*. Counting less frequent conventional variants as variation would overestimate the degree to which an idiom is varied, as it is not productive variation. It would be interesting to see how much productive lexical variation of this type there is. But this would be very hard to study because there would be huge amounts of irrelevant data to sift through as one would have to look at every sentence containing each of the words occurring in the idiom independently, which can be a very large set of data for frequent words.[4] In any case, this would be a very different type of variation than the one I am talking about and proposing an approach to.

The type of variation 'internal to' the combination of a particular verb with a particular idiomatic complement can be classified into several different types. One common type is changing a noun phrase which is part of the idiom, e.g. by changing the specifier, adding a modifier, or changing the number of the noun. Other variations are more syntactic in nature and include passivization and occurrence across relative clause boundaries.[5]

[3]For full citations of all dictionaries used see Appendix B.

[4]For some idioms containing rare words I searched just for those words, and was therefore able to see what contexts they occur in.

[5]The use of the term 'variation' in this dissertation should not be confused with its use in

I assume that there is no amount of variation that would make an occurrence not be an instance of the idiom, as long as the idiomatic meaning is present (although of course there may be some individual corpus examples where the sentential context makes it is hard to determine this with certainty). That is, I classified every utterance containing both a form of the word *spill* meaning 'reveal' and a form of the word *beans* meaning 'secrets' as an instance of the idiom *spill the beans*.[6]

The definition of 'canonical form' is a bit tricky. It is essentially the form in which the idiom occurs most frequently in the data, modulo inflection of the head (if there is one). For example the canonical form of the idiom *spill the beans* is *spill/spills/spilled/spilling the beans*, while *spilling many beans*, *spilled the royal beans*, and *the beans were spilled* are non-canonical occurrences. However, defining 'canonical form' this way would be circular, as this form would by definition be the most frequent one.[7] So I instead define 'canonical form' as the main form listed in the Collins COBUILD idioms dictionary. This form almost always corresponds to the form listed in the NTC idioms dictionary and the most frequent form occurring in the data. Of the 22 idioms in the random sample from the Collins COBUILD idioms dictionary, 9 are not listed in the NTC idioms dictionary.[8] Of the 11 idioms that are listed in both dictionaries, 9 are listed in identical form, although the COBUILD dictionary also mentions variants of the idioms in some cases. The other two idioms are *take a back seat*, which is listed as *take a backseat* in the NTC dictionary, and *rear its (ugly) head*, which is listed as *rear its ugly head* in the NTC dictionary while the COBUILD dictionary lists both *rear its head* and *rear its ugly head*. As will be argued below, *take a backseat* is just a spelling variant, and *rear its ugly head* should

sociolinguistics. The term was chosen because 'variant' is conventionally used in idioms dictionaries to refer to one of the conventionalized variants, which usually involves substituting different lexical items.

[6]I included in the data a few examples where idiomatic words are coordinated with non-idiomatic ones or take an idiomatic and a non-idiomatic complement, although it is not totally clear whether the words have their idiomatic meaning in these examples. Most of these examples should probably be called 'word play'.

[7]Note that it would still be significant that a particular form of an idiom is so much more frequent than would be expected by comparison with a semantically and syntactically similar non-idiomatic expression.

[8]These are: *deliver the goods*, *run the show*, *pay dividends*, *sound the death knell*, *break the mold*, *level the playing field*, *lead the field*, *hit home*, and *speak volumes*.

be considered an established variant of the idiom and not productive variation. The two dictionaries do not disagree about issues like number or definiteness of the complement. In all cases the main form listed in the COBUILD dictionary is also the most frequent form occurring in the corpus, so for this sample this definition is in 100% agreement with the characterization of 'canonical form' as the most frequently used one.

It is not completely clear whether examples involving relative clauses modifying the idiomatic NP, such as *level the playing field that IBM once dominated* should be considered canonical. In these examples the string *level the playing field* is present as in the canonical form, and for decomposable idioms these examples do not pose a problem to most approaches, as the relative clause does not interfere with the relationship between *level* and *playing field*. They only pose a problem to approaches that represent the idiom as a complete VP. However, these examples of modification provide one piece of evidence that an idiomatic noun can have meaning associated with it.

The next question is how to determine whether an idiom is semantically decomposable, i.e. whether the parts of the idiomatic meaning are associated with parts of the idiom. It is not the case that for every idiom speakers will agree on what its precise meaning is, and in some cases it will be a decomposable meaning for some speakers and a non-decomposable one for others. Another issue is that there are competing constraints: speakers want preassembled chunks that can just be used, but they also want to analyze things and to make sense of them. So it is by no means straightforward to classify all idioms correctly, and it is not totally clear that there are only two clearly separable categories.

The least biased test for decomposability is the degree of variation an idiom shows, as this is in some sense a poll of how other speakers perceive the idiom.[9] But obviously this is completely circular: if I define every idiom with a very low degree of variation as 'non-decomposable', then it is not at all surprising that non-decomposable idioms have a very low degree of variation. Fortunately my best effort at coming up with a

[9]As was shown in Chapter 2 certain types of variation, such as internal modification and passivization are possible only for decomposable idioms.

decomposable interpretation tends to coincide with the corpus' verdict.[10] Note that there are two steps to the process of trying to find a decomposable interpretation. The first is to come up with a meaning or paraphrase with the same semantic structure as the idiom. But even when that can be found, there is still the question of whether the idiomatic noun can be thought of as carrying part of that meaning. For example, while *clear the air* can mean 'eliminate the misgivings', for most speakers it is not the case that *the air* directly corresponds to 'the misgivings'. Instead 'the misgivings' are usually thought of as what is being 'eliminated' from the *air*, which corresponds to the situation as a whole. But obviously there is no guarantee that every speaker sees this the same way, and sometimes it is possible to see from varied corpus examples that some speakers have analyzed an idiom differently from other speakers.[11]

Given that the parts of semantically decomposable idioms carry parts of the idiomatic meaning, it is to be expected that they can be internally modified (Nunberg et al. 1994). For example, since *spill the beans* can be analyzed such that *beans* plays the role of 'secrets' in the meaning of the idiom, it is not surprising that *beans* can be modified to further specify what kind of secrets are being revealed, as in *Diana spilled the royal beans*.

Non-decomposable idioms are not expected to show as much variation because their parts cannot be associated with parts of the meaning of the idiom. One kind of variation that is observed is the meta-linguistic type using adjectives like *he kicked the proverbial bucket*. Another type can be observed in the example *they kicked their respective buckets*. In both types, the adjective modifies part of the idiom syntactically but not semantically. This is called external modification, as the scope of the adjective

[10]The one exception is the idiom *take a back seat*, which I classified as decomposable, but which shows an unusually low degree of variation for a decomposable idiom.

[11]Some unclear cases are discussed in Section 3.5. Given that speakers may differ in how they analyze a particular idiom, it is theoretically possible that speakers with a non-decomposable analysis never vary it, and speakers with a decomposable analysis vary it as frequently as non-idioms. But this does not seem to be the case. While I do not know how to get a large enough sample of individual authors' writing to really test this claim, a search in Lexis-Nexis found some newspaper writers who vary idioms some of the time but still mostly use the canonical form. Leah Garchik varied 7 out of 22 occurrences of *spill the beans*, Stephen Holden varied 2 out of 7 occurrences of *turn the tables*, and William Safire varied 2 out of 5 occurrences of *spill the beans*, 3 out of 8 occurrences of *turn the tables*, and 4 out of 10 occurrences of *make waves*.

is external to the idiomatic VP.

3.2 Idioms Discussed in the Literature

All the idioms in this section were discussed by Nunberg et al. (1994), except where otherwise indicated.

3.2.1 Decomposable Idioms

spill the beans

87% of the occurrences of this idiom are in their canonical form: 97 out of 111 show no variation other than inflection of *spill*.
There are 14 non-canonical examples of this idiom, which represents 13% of the occurrences in the corpus.

- In 7 examples there is a different specifier (*some beans*, *all the beans*, *any beans*, *mountains of beans*, *whatever beans he manages to find*)
- 5 examples have a modifier (*royal beans*, *some very complimentary beans*, *the politically charged beans*, *the Boston baked beans*, *the oboistic beans*)
- 2 examples involve compound nouns (*the jelly beans*, *the Arkansas beans*)

These 'quantified' and 'modified' occurrences of *beans* sound quite natural in context:

(69) a. No soap opera worth its bubbles would spill all the beans in one episode if it could dribble them out over many.

b. The courtroom climax is rigged in more ways than one, with the victim put implausibly on trial and a key witness spilling mountains of beans conveniently on cue.

c. Diana to Spill Royal Beans in Upcoming Interview

d. Yet more and more of Bill and Hillary Rodham Clinton's time is being spent thinking about special prosecutors, angry FBI directors and old friends who might spill the Arkansas beans for reduced sentences.

Note that (69a) also contains a pronominal reference to part of the idiom.

pull (the) strings

The study of this idiom is complicated by the fact that there are two very similar idioms with different although related meanings: *pull the strings*, meaning 'being in control', vs. *pull strings*, meaning 'exploit connections'.[12] Here are two typical uses of the idiom *pull strings*, meaning 'exploit connections':

(70) a. The architect on the project pulled strings with city officials, who awarded a grant for exterior restoration.

b. He spent a year pulling strings to get the script to a large selection of well-known actresses nearly 40 and older.

Some typical uses of the idiom *pull the strings*, meaning 'being in control' are:

(71) a. [...] Karadzic still pulls the strings from behind the scenes [...]

b. He's the guy who's really pulling the strings in the mayor's office [...]

c. Even from jail, he seemed to continue to pull the strings of the underworld in Marseille [...]

However, the meaning of the two idioms is obviously related—in most cases 'exploiting connections' also involves 'exercising control' or 'exerting influence'.[13] This and insufficient context made it very hard to classify the corpus examples into uses of the two idioms. A further complication comes from the fact that many uses of *strings* invoke the metaphor more directly—many examples involve *puppet* or *puppeteer* (e.g. *the CIA were pulling the strings and Suu Kyi was acting as their puppet*, *Nancy Reagan as the puppeteer who pulled the strings of her husband's presidency*, and *...pulled premiers' strings like a master puppeteer*). These examples cannot be understood using the paraphrases *exercise control* or *exert influence*. While examples

[12]This is the meaning that Nunberg et al. (1994) give for the idiom *pull strings*. The two idioms also have separate entries in the Collins Idioms dictionary, where *pull strings* is defined as 'get something one wants by using one's friendship with powerful and influential people', and *pull the strings* is defined as 'control everything that another person or organization does'. The NTC dictionary only lists one of these idioms, *pull strings*, which they define as 'to use influence (with someone to get something done)'.

[13]But one can at least try to 'exploit connections' without having any influence or control over the other person, and it is fine to say *I went to High School with Bill Clinton, so I tried to pull strings with him*, while ??*I went to High School with Bill Clinton, so I tried to exert influence/exercise control over him* is somewhat strange and might even suggest attempted blackmail.

involving these words can be identified quite easily, other examples are not so clear. Many examples involve *stage*, *scene* or *the strings of NP*, where an explicit analogy to a puppet is not necessarily present, but may well be intended. Note also that if *pull strings* is supposed to mean *exert influence* and not *exploit connections*, then examples involving *pull strings with* as in (70a) cannot be interpreted compositionally, either, and one would need a separate representation for *pull strings with* corresponding to *exert influence over*. For these reasons it is quite impossible to classify the 266 examples into the two idioms and the metaphorical uses of *strings*, and I will not attempt to do so.[14] But the more unusual type of variation that occurs is still interesting, so it is presented below.

Some passivized examples are given in (72):

(72) a. A week later, a suitable donor was found. No strings were pulled.
b. Even as savvy a political operator as D'Amato has only theories on what political strings may have been pulled.
c. [...] the small American Communist Party was little more than a propaganda and espionage tool whose strings were pulled by the Kremlin.
d. She says the process will empower voters who understand that the strings are being pulled by somebody.

Some examples are unusual in other ways, such as occurring across clause boundaries:

(73) a. Why would a reputed master politician fail to find one string to pull?
b. At least Pat Robertson was in harm's way long enough to have strings pulled to get him out.
c. [...] Robert McNamara's new book justified all the strings Clinton pulled as a young man to keep his precious hide out of harm's way in Vietnam.
d. Not even all the strings that Sounders coach Alan Hinton pulled to get Baggio to make an appearance was enough to fill the stadium.
e. They had the strings and pulled them.
f. But this typed expression of thanks for strings pulled was slightly different.

[14]Note that there are 74 examples of the form *pull strings* and 81 examples of the form *pull the strings*, so a rough estimate suggests that each of these idioms is over 50% canonical.

g. But Lee has even more strings to pull [...]

h. The only strings that will be attached are those that you as taxpayers get to pull when you decide every year how your money can best be spent by local officials

Note that (73h) uses *strings* in two apparently unrelated idiomatic meanings at the same time. This suggests that the speaker thought of *strings* as having the same meaning in both cases. If the speaker was trying to invoke multiple meanings, this example is unacceptable for the same reason as **the river will flow over the bank that you have your account at*. Indeed, (73h) is unacceptable for many speakers, who are presumably unable to think of a meaning for idiomatic *strings* that fits both idioms.

keep tabs on

This idiom is discussed for example in Bresnan (1982:49), Gazdar et al. (1985:236), and Nunberg et al. (1994:502). 72% of the occurrences of this idiom are in their canonical form: 418 out of 578 show no variation other than inflection of *keep*.
There are 160 non-canonical examples of this idiom, which represents 28% of the occurrences in the corpus.[15]

- 16 examples do not include *on*. 5 of these are also modified (*keep close tabs* (4), *keep better tabs*). 5 of them are of the form *keep tabs of*
- In 6 examples *tab* is singular, and 4 of these are also modified (*a close tab* (2), *a pretty close tab*, *a continuous tab*)
- 130 examples have a modifier (*close* (77), *closer* (24), *such close*, *very close* (3), *fairly close*, *unusually close*, *surprisingly close*, *better* (6), *regular* (3), *strict* (2), *smitten*, *monthly*, *surreptitious*, *exact*, *daily*, *careful*, *moment-by-moment*, *constant*, *secret*, *detailed*, *loose*)

5 examples involve an *on* complement, but it is not adjacent to *tabs*:

(74) a. [...] she kept tabs by telephone on the theater

[15] If *keep close tabs on* is considered a conventionalized variant of this idiom then modification by *close* should not be considered productive variation. If the 119 occurrences involving *close* are counted separately, 77 (65%) of them are canonical. Of the remaining 459 occurrences of this idiom, 91% are canonical.

b. [...] promised to keep tabs as well on the official police investigation.

c. The Guatemalan army keeps close tabs not just on suspected guerrilla supporters, but on a wide variety of academics [...]

d. The report says White House lawyers kept close tabs not only on the RTC's work but on an investigation by the Small Business Administration of David Hale [...]

e. Flight directors were keeping close tabs Tuesday on a frontal system in the Midwest.

1 example goes across a relative clause boundary:

(75) But the governor's biggest contribution may be the close tabs he keeps on his ground operation.

2 examples involve raising or pied-piping:

(76) a. But Baucus, who supported the reintroduction of wolves, wants better tabs kept on the wolves.

b. [...] guerrilla and other groups on which the United States wants to keep tabs

pay the piper

77% of the occurrences of this idiom are in their canonical form: 33 out of 43 show no variation other than inflection of *pay*.

There are 10 non-canonical examples of this idiom, which represents 23% of the occurrences in the corpus.

- In 1 example *piper* is modified by an adjective (*public piper*)
- 8 examples are passive, and 6 some of them also involve raising auxiliaries (*the piper is paid*, *the piper must be paid* (3), *the piper will be paid*, *the piper is to be paid*, *the piper may soon have to be paid*, *there is a piper to be paid some time*)
- 1 example is *the San Francisco 49ers may have a demanding piper to pay next year*

bury the hatchet

This idiom is discussed by Jackendoff (1997). 76% of the occurrences of this idiom are in their canonical form: 112 out of 147 show no variation other than inflection of *bury*.

There are 35 non-canonical examples of this idiom, which represents 24% of the occurrences in the corpus.

- In 12 headline-style examples there is no specifier (*bury hatchet*)
- 1 example has a different specifier (*Hannah's hatchet*)
- 5 examples have a modifier (*ideological*, *legal*, *partisan*, *ethnic*, *historical*)
- 1 example involves a compound noun (*the family hatchet*)
- 1 example is modified by a relative clause (*bury the hatchet that has been swinging around on this island for the last 25 years*)
- In 7 examples *hatchets* is plural. All of these are modified in some way (*their hatchets* (2), *any hatchets*, *their takeover hatchets*, *the hatchets of old blood feuds*, *one of the sharpest hatchets left in the post-Cold War world*, *there are some very sharp hatchets to bury*)
- 1 example is *a burying of the hatchet*

7 examples are passive and one of them also involves raising:

(77) a. Now that the hatchet has been buried, they can once again focus on their business
b. The hatchet is buried.
c. For the families of Wilde and Douglas, the hatchet has long been buried.
d. The hatchet has since been buried, the officials say.
e. [...] the hatchet was finally buried in 1991 after the soccer union's launch of a huge anti-hooliganism campaign.
f. The Wilde and Queensberry families feuded bitterly at the beginning of the century, but the hatchet is now truly buried.
g. The hatchet now appears to have been buried for good [...]

raise hell

I studied this idiom because it was discussed by Jackendoff (1997:170), who said that it does not readily undergo passive even though it has a plausible semantic decomposition ('cause a serious disturbance'), and that it should be analyzed as a lexical VP rather than a collection of syntactic fragments. The corpus data show that this is not the case.

80% of the occurrences of this idiom are in their canonical form: 98 out of 123 show no variation other than inflection of *raise*.

There are 25 non-canonical examples of this idiom, which represents 20% of the occurrences in the corpus.

7 examples are of the form *raise holy hell*. I am not sure whether this is a conventionalized variant of this idiom, as it is not listed in my idioms dictionaries. If it is a conventional variant, these examples should be disregarded, and in that case the percentage of variation for this idiom is 84%.

9 examples are modified: (*some hell* (2), *too much hell*, *a little hell*, *enough hell*, *more hell*, *the highest kind of hell*, *absolute legislative hell*, *raising the kind of hell we raised in better days*)

1 example is passivized:

(78) So much hell was raised that the biologists threw up their hands in surrender

3 examples occur across relative clause boundaries, and one of them is also passivized:

(79) a. [. . .] the internal investigation was reopened "in part because of the hell that Plitman raised about Newcomb's role in Leatherneck."

b. What we're hoping is, the amount of hell we can raise in the United States and the heat that G.E. takes will make them cooperate

c. Few folks in the Apple speculated on the hell that would have been raised by George Steinbrenner if the Yankees had been similarly robbed at Camden Yards.

5 examples are varied in other ways:

(80) a. Sure there is night life, but how much hell can you raise at Jack-in-the-Box?

b. The more hell they raise, the better Clinton looks

c. This will be heck raising. But not, he insists, hell raising.

d. But all they really raised was a little hell and a lot of dust.

e. But like 19th-Century Populist Mary Elizabeth Lease, who exhorted Kansas farmers to "raise less corn and more hell," they do get riled up with great regularity.

the cat ... out of the bag

This idiom is discussed e.g. by Pulman (1993). It turns out that *let* is not an essential part of this idiom—23 out of 48 occurrences of the idiom in the corpus do not contain *let*. It is not clear how to define 'canonical form' for such a (possibly discontinuous) idiom, and I will not attempt to do so.[16]

- 15 examples are of the form *let the cat out of the bag*
- 14 examples are of the form *the cat is/was out of the bag*
- 3 examples involve event modification (*the cat was really out of the bag*, *the cat was already out of the bag*, *this cat is 99 percent out of the bag*)
- 7 examples involve changing the specifier of 'cat' and/or adding a modifier or a compound noun (*Schrodinger's cat*, *too much of the cat*, *the angry cat*, *the big cat*, *that little cat*, *the tax-cat*, *the inflation cat*)
- 1 example adds a modifier for 'bag' (*the dynastic bag*)
- 2 examples involve both 'cat' and 'bag' modification (*let the strategic cat out of the fiscal bag*, *let the neon cat out of the cellophane bag*)

Some full examples are given below:

(81) a. Once it made the cover of Entertainment Weekly in March, the cat was really out of the bag.

b. Babbitt's real mistake was to let the tax-cat out of the bag before Election Day.

c. Daschle didn't want to let too much of the cat out of the bag before the formal announcement [...]

[16] One possible approach would be to treat it as having multiple established forms.

d. But the Senate Finance Committee let the strategic cat out of the fiscal bag when it released a new study [. . .]

e. OK to reveal that the playoffs are the tail that wags the NBA regular season dog is to let the neon cat out of the cellophane bag.

Note that the interpretation (81e) requires reference to the literal meanings (a 'cellophane bag' is something transparent), and has to be understood metaphorically.
Other variations include:

(82) a. The cat got out of the bag this month [. . .]

b. It looks as if Whitmore tried, unsuccessfully, to stuff the cat back in the bag.

3.2.2 Non-Decomposable Idioms

kick the bucket

11 out of 12 occurrences of this idiom (with the meaning 'die') are in their canonical form (92%).
The one non-canonical occurrence of this idiom is:

(83) We're initially drawn into them by the discovery of corpses and the question of who made them kick their respective buckets.

Note that this is not semantically internal modification, i.e. it means 'who made them die, respectively'.

saw logs

There are only 3 occurrences of this idiom (with the meaning 'snore') in the corpus. It is interesting although not statistically significant that they are all varied, as can be seen in (84).

(84) a. A friend of mine whose husband has been sawing some major logs every night of their marriage has tried everything.

b. I still sawed a few logs, or so alleged my wife, Jane.

c. [. . .] Charlton was in the clubhouse, sawing major logs.

The meaning of *major* could be external modification, i.e. *snore in a major way*. But *a few* seems to be an example of internal modification. Perhaps for this speaker the idiom is actually decomposable and means something like *make snoring noises*, although this does not match the role of *logs* in the metaphor. It is more likely that *a few* was used in the context of the metaphor.

shoot the breeze

47 out of the 51 occurrences of this idiom (with the meaning 'chat') are in their canonical form (92%). The non-canonical occurrences of this idiom are:

(85) a. First up on this night was Ahmad Rashad [...] who dropped into the seat next to Lee with the game more than an hour away, to shoot the hoops breeze [...]

b. In the Senate Chris Dodd, Bill Bradley, John Kerry, John McCain and Al D'Amato love to shoot the on-air breeze with the craggy-faced shock jock.

c. The earliest citation in this sense of shooting pyrotechnic breeze is in a 1967 article in the magazine Trans Action.

d. I carried my gun at the ready, safety lock on, but I had no desire to shoot anything more than the breeze.

The last example should probably be called word play because it involves a pun exploiting two different meanings of *shoot*.

make tracks

All 10 occurrences of this idiom (with the meaning 'leave') are in their canonical form. An example is given in (86).

(86) [...] she'd made tracks out of her native Oklahoma [...]

take a powder

All 18 occurrences of this idiom (with the meaning 'leave') are in their canonical form. An example is given in (87).

(87) ABC's "NYPD Blue" could have dropped out and its fans tuned out when David Caruso took a powder.

hit the ceiling

16 out of 17 occurrences of this idiom (with the meaning 'get angry') are in their canonical form. A typical canonical occurrence of this idiom can be seen in (88).

(88) She would like to cut his $300-a-month allowance to zero, but he would hit the ceiling if she did.

The non-canonical occurrence of this idiom is:

(89) Even the stoic Encyclopedia Britannica hit the proverbial ceiling over the unsatisfactory quality of ceilings today [...]

Note that this idiom should not be confused with the expression *hit a (glass) ceiling* meaning 'reach a limit'.

hit the roof

All 19 occurrences of this idiom (with the meaning 'get angry') are in their canonical form. A typical canonical occurrence of this idiom can be seen in (90).

(90) Cerniglia said he hit the roof when he found out his son was giving some of his pocket money to a politician.

3.2.3 Summary

Table 3.1 summarizes the results of the study of idioms from the literature.

3.3 Idioms Relevant for Argumentation

In this section I study some idioms that have interesting properties, such as occurring mostly in passivized form, occurring mostly in negated form, containing adjuncts, not involving a verbal head, and containing more than one idiomatic noun.

	Total # of Tokens	% Canonical	% Variation
Decomposable Idioms			
spill the beans	111	87%	13%
pay the piper	43	77%	23%
bury the hatchet	147	76%	24%
raise hell	123	80%	20%
Average:		**80%**	20%
Non-Decomposable Idioms			
kick the bucket	12	92%	8%
shoot the breeze	51	92%	8%
make tracks	10	100%	0%
take a powder	18	100%	0%
hit the ceiling	17	94%	6%
hit the roof	19	100%	0%
Average:		**96%**	4%

Table 3.1: Results from the Study of Idioms from the Literature

caught short

81% of the occurrences of this idiom, meaning 'caught unawares' are in their canonical form: 84 out of 104 are passives of the form *caught short*.

1 example is passive but varied:

(91) The airlines could be caught very short

6 examples are passives of the form *caught up short*.

13 examples are active (12 of them involve the past tense form *caught*):

(92) a. [...] the new world corporate order has caught the Russians short [...]

b. [...] the United States caught its trading partners short by proposing a possible compromise [...]

c. But the old candidate expects crime, patriotism and race to not catch the president short.

Note that I excluded all clear occurrences of the 'insufficiently supplied' meaning of *short* (which often takes the form *short of NP*) and the stock market meaning of *short*, although it is sometimes unclear which meaning is intended. However, these other expressions occur passivized to a similar extent when they collocate with *catch*.

be laughed out of court

10 out of 10 corpus examples of *be laughed out of court* are passive.

caught in the middle

138 out of 139 corpus examples are of the form *caught in the middle*. There are no examples of the form *catch/catches/catching NP in the middle*, and *caught* is clearly not a past tense form in any of these 138 examples. It is consistent with being an adjectival passive with a stative interpretation in all the cases, although this cannot always be clearly distinguished from a verbal passive interpretation.
Some fairly clear cases of adjectival passive participle uses with a stative interpretation are given below:

(93) a. Trainers feel caught in the middle of the labor battle.
b. Hundreds of foreign travelers found themselves caught in the middle of the chaos on Wednesday [...]
c. If you are a taxpayer caught in the middle, the situation can be frustrating.

However, most examples involve a form of the verb *be*, and in those examples and in some other cases one cannot completely rule out a verbal passive interpretation, although there is no example of an overt *by* phrase in the corpus.

(94) a. The civilians are caught in the middle [...]
b. He is caught in the middle of what has become a nasty situation between the Giants and Jarrod Bunch [...]
c. Sponsors tend to get caught in the middle of these disputes because they have money and reputations at stake.

The one clearly non-canonical, i.e. active example is:

(95) [...] a bombshell that exploded over the industry last week and caught investors in the middle.

For most speakers there are probably strong constraints on when this idiom can be used actively. Note that in this example it is coordinated with another VP, and it is in the past tense. And some speakers find even (95) ungrammatical.

not born yesterday

For idioms where the negated form is the conventionalized one it is not obvious how to define 'canonical form'. Is it any occurrence of *not*, does it matter whether it is full or contracted, and does it have to be adjacent to the rest of the idiom? It is not obvious what the right answer is, and it might be different from speaker to speaker what their representation for the canonical form of such idioms is, as well as what types of variation they accept. It would be necessary to have a larger set of examples and some judgment data to decide this.[17]

19 out of 26 occurrences of this idiom are instances of negated past tense forms of *be* (13 *wasn't*, 2 *was not*, 4 *weren't*).

The other examples also involve some sort of negative interpretation, but only some of them are standard negative polarity contexts.

(96) a. Well, I am not a consumer born yesterday.
b. None of us was born yesterday.
c. Mandela does not want to be seen by his peers as the child who was born yesterday who is preaching to the elders
d. The Yiddish language and its literature were hardly born yesterday.
e. Excuse me, do I look like I was born yesterday?
f. They must think I was born yesterday.
g. The Republicans are behaving as though they think we older folks were born yesterday.

(*to*) *put it mildly*

76% of the occurrences of this idiom are in their canonical form: 80 out of 105 occurrences[18] are of the form *to put it mildly*, as in (97a).[19]

19 examples are of the form (*that*) *is/was/be putting it mildly* as in (97b) and (97c).[20]

[17] A separate question is how these types of constraints can be expressed in each particular theory of idioms, and it seems that phrasal approaches are at an advantage here.

[18] This search was conducted only in the NYT part of the corpus.

[19] It is somewhat unclear whether *to* should be considered part of the canonical form. If it is not, then 94% of the occurrences of this idiom are canonical.

[20] Perhaps *putting it mildly* should be considered a second established form of this idiom.

6 examples show other variations, as in (97d) and (97e).

(97) a. It is a can of worms, to put it mildly [...]
b. Stan Belinda and closer Heathcliff Slocomb, in particular, were shaky – and that's putting it mildly.
c. Aggressive is putting it mildly.
d. No, that's putting it too mildly.
e. That's putting it more than mildly.

One might think that *put it mildly* is a collocation and not an idiom, because the verb *put* can occur with a fairly similar meaning in other places. However, it usually requires two complements (e.g. *she put her thoughts into words*). Furthermore, in the expression *to put it mildly*, *it* is obligatory, and #*he put his objection mildly* sounds strange. In addition, *mildly* is an odd choice of adverb: #*your way of putting this is mild* sounds strange. If *mildly* could have exactly the same meaning outside of this idiom, one would expect at least some corpus examples like *to say it mildly*, *to express it mildly*, or *to state it mildly*, but those are not found. Furthermore, even if they were found, I suspect the statements they refer to would have to be more literally *mild*, i.e. 'moderate' and not 'harsh'. In contrast, the idiom *to put it mildly* does not have go with 'moderate' statements. It merely suggests that the author would have liked to use an even harsher statement. That is, this idiom marks understatements. This can be seen when one examines some of the statements that it is used with, e.g. *furious*, *ludicrous*, and *pathetic*. In some cases it is hard to see what harsher statement could have been used.

(98) a. well, to put it mildly, the Coyotes have *blown it*
b. It was *hokey, hackneyed and horrible*, to put it mildly [...].
c. [...] others are *furious* about his insensitivity, to put it mildly.
d. The government's reaction to the Freemen has been, to put it mildly, *pathetic*.
e. [...] the idea [...] is, to put it mildly, *ludicrous*.
f. The dollar is *plunging* to put it mildly [...]

Furthermore, only 2 examples are of the form in (99), i.e. talking about an agentive subject actually saying something. One might expect this form to be more frequent if this was a collocation meaning something like 'say it in a nice way'.

(99) Barbara Boxer *put it mildly*, "There is a certain irony here."

fall on deaf ears

90% of the occurrences of this idiom are in their canonical form: 252 out of 281 are of the form *fall on deaf ears* modulo inflection of *fall*.
29 examples are noncanonical.

- 5 examples are of the form *fall upon deaf ears*
- 1 example is *fell on the deaf ears of most of the other candidates*
- 10 examples are of the form *fall ADV on deaf ears* (*fall largely on deaf ears* (5), *fall mostly on deaf ears* (3), *fall increasingly on deaf ears*, *fall mainly on deaf ears*)
- 13 examples are of the form *fall on MOD deaf ears* (*fall on many deaf ears* (3), *fall on some deaf ears* (3), *fall on even more deaf ears*, *fall on mostly deaf ears*, *fall on apparently deaf ears*, *fall on seemingly deaf ears*, *fall on altogether deaf ears*, *fall on official deaf ears*, *fall on delicate but deaf ears*)

hide one's light under a bushel

7 out of 10 of the occurrences of this idiom are in their canonical form, i.e. *hide one's light under a bushel*. The remaining 3 examples are:

(100) a. Its principal architect is Kenneth Branagh, who does not hide his light under bushels
b. What writer hides his or her light under the bushel of anonymity these days?
c. It's quite feasible to argue that New Holland's light is being hidden behind the bushel in the Fiat context

needle in a haystack

I thought *look for a needle in a haystack* was an interesting idiom to study because *in a haystack* is not a complement of *look*. However, it turns out that there are only 9 examples of *look for a needle in a haystack* that are not of the form *like looking for a needle in a haystack* (which is not technically an idiom but a simile, so that these 10 occurrences are excluded from the data).[21] Furthermore, there are 85 examples of *needle in a haystack* that do not involve *look*, so that *needle in a haystack* has to be considered the actual idiomatic expression. This makes the variation data for this idiom quite different in nature, but it is still interesting.

- 52 out of 94 examples are *needle in a haystack* (including 5 examples of the form *needle-in-a-haystack*)
- 14 examples are *needle in the haystack* (including 1 needle-in-the-haystack)
- 4 examples are *needles in a/the haystack*
- 3 examples involve a modified 'haystack' (*in a cosmic haystack*, *in a great big haystack*)
- 1 example changes the 'needle' to *8 3/8 million needles*
- 7 occurrences modify both the 'needle' and the 'haystack' (*an information needle in the mammoth haystack of the Web*, *a 6-foot needle in a mountainous haystack*, *a deadly needle in a giant haystack*, *the troublesome needles in the Palestinian haystack*, *that fetal needle in the maternal haystack*, *that needle of information in the electronic haystack of pages out there*)

Other variations include:

(101) a. Take a look at the haystacks again; I think there might be needles in them.

b. Over the years, Cullers has helped develop ways to use computers to probe through dense "haystacks" of space signals in hopes of finding the one precious "needle": an intelligent message from another world.

c. One has to blow away the climate haystack in order to find a couple of very, very small needles.

[21] Of the 9 examples of the collocation *look for a needle in a haystack* 4 are canonical.

d. The hunt for a China policy that is at once moral, profitable and politically tenable increasingly resembles the search of a haystack that does exist for a needle that doesn't.
e. The whole transit zone north of Colombia is the biggest haystack in the world with diverse, different types needles all around.
f. Mr. Simpson's lawyers will use advanced searching software to quickly identify all the relevant documents, much like running a magnet through a haystack to extract all the needles.
g. Yet, in the midst of this haystack of data, certain needles may be found.
h. "They might as well throw a needle onto a haystack and try to sell that" because the Web is so vast and complex.
i. Such needle-in-haystack searches have provided an extra dose of drama.
j. And he ridiculed some of the findings as "haystacks without needles".

separate the wheat from the chaff

71% of the occurrences of this idiom are in their canonical form: 20 out of 28 are of the form *separate the wheat from the chaff*.
There are 8 non-canonical examples of this idiom. In 2 examples there is no specifier for both nouns (*separate wheat from chaff*). In 6 examples one or both of the nouns in the idiom are modified:

(102) a. And the lawsuits, allegations, losses and stricter oversight have separated much of the wheat from the chaff at Lloyd's.
b. Here's a rundown separating the video wheat from the chaff.
c. New Hampshire voters have spoken, but not authoritatively, as they tried to separate the political wheat from the chaff but left only a few percentage points between candidates.
d. He has assigned three agents to the property and set up an elaborate system to separate the wealthy home-buyer wheat from The Trial-obsessed chaff.
e. Or has the East found a way to separate the wheat of economic growth from the chaff of Western "decadence"?

1 example is also passive:

(103) Next year will be the year the information wheat is separated from the digital chaff

3.4 Idioms Containing Non-Independent Words

I also studied two idioms containing words that never occur outside of the idiom: *tit for tat* and *by dint of*. The noun *tat* is not listed in Merriam-Webster's Collegiate Dictionary and never occurs in the corpus by itself, although it is a word in British English. The noun *dint* never occurs in the corpus by itself. The NTC idioms dictionary says: "*Dint* is an old word meaning 'force' and it is never used except in this phrase." These expressions meet my definition of idiom because words that never occur by themselves necessarily never occur by themselves with the same meaning.

However, it would be better to find idioms of this type that also contain a verb. There are some idioms like *play hooky* and *run roughshod over* containing words that hardly occur outside of the idiom and do not seem to be independent words in the mental lexicon of some speakers. These speakers accept examples containing the whole idiom, but not ones where words like *hooky* and *roughshod* occur independently. Some of these speakers also accept varied examples like *play a bit of hooky* and *run blithely roughshod over*. Other promising candidates occurred more frequently than I expected: *hotcakes* occurred 15% of the time outside *selling like hotcakes*, *edgewise* occurred 13% of the time outside of *(not) get a word in edgewise*, and *wedlock* occurred 5% of the time outside of *(born) out of wedlock*. I include the data from these studies anyway, as there may be speakers for whom these are truly non-independent words.

by dint of

112 out of 125 occurrences of this idiom are in their canonical form (90%).
The idiom is clearly non-decomposable for most speakers, for whom it is like *because of* or *due to*, but there seem to be some speakers for whom the idiom is decomposable, meaning something like *by force of*. Of the 13 non-canonical occurrences of this idiom 4 occurrences are of a type that does not necessarily involve semantic modification of

the parts of this idiom: *by mere dint of*, *by pure dint of*, *by dint not of*, and *by dint only of*. These mean *merely because of*, *purely because of*, *not because of*, and *only because of*, respectively. But 6 occurrences are *by sheer dint of*, which does not seem consistent with a non-decomposable meaning (?*sheerly because of*).[22] 3 occurrences are actually *through dint of*, suggesting that for some speakers *dint* might still have a separate lexical entry. In any case, decomposable or not, this idiom clearly shows that even this idiom cannot be given an analysis treating it simply as a lexicalized complete phrase.

tit for tat

The idiom *tit for tat* is canonical 95% of the time and varied in two corpus examples:

(104) a. For every tat you get a double tit.

b. These exercises were their tit to our tat.

The type of variation in the last example either involves creative word play, or this idiom is decomposable to some speakers—perhaps speakers of British English, where *tat* is an independent word.

play hooky

The word *hooky* meaning *truant* or *truancy* occurs by itself 11% of the time (in 8 examples), so it is an independent word for some speakers. However, an informal native-speaker survey turned up several speakers who accept occurrences of idiomatic *play hooky* but reject the corpus examples containing *hooky* by itself:

(105) a. It's a shame they'd rather declare war on hooky.

b. The track is not exactly hooky headquarters.

So at least for some speakers this idiom contains a word that does not seem to exist independently.

[22] However *sheerly* is a very infrequently used adverb—there are only 4 occurrences of it in the 350 million words of the corpus. So it is not surprising that some speakers do not like it as an adverb at all, and others accept it only in some constructions but not others. In fact, there is at least one speaker who accepts *sheerly because of*.

There is only one non-canonical occurrence of idiom *play hooky*, which has its canonical form 98% of the time:

(106) To the chagrin of his speech-writing team, the president has played a bit of hooky during his rail trip to Chicago, as he took time to mingle with admirers and hail onlookers from the rear platform of the train.

There is no way to know whether the word *hooky* has an independent life for the person who wrote this sentence, but it is acceptable for at least some of the speakers who reject occurrences of *hooky* without *play*.

run/ride roughshod over

The word *roughshod* occurs outside of the expression *run/ride roughshod over* 13% of the time. If one applies the more generous criterion that it is sufficient for either *run* or *ride* or *over* to be present, it still occurs outside of these contexts 4% of the time. So it is clear that it is an independent word for some speakers. However, a few speakers reject the corpus examples containing *roughshod* by itself:

(107) a. Ev & El's venture was more roughshod than Ms. Kahng's.
b. Here are a few other examples of roughshod justice.

The idiom *run/ride roughshod over* occurs in one of its two forms *run roughshod over* and *ride roughshod over* 89% of the time. Variations include *run absolutely roughshod over*, *run blithely roughshod over*, *run over everybody roughshod*, and *running roughshod not only over*. But there is also further variation in the choice of verb: *gallops roughshod over*, and *driving roughshod over*, and in the choice of preposition *ride roughshod across*, *run/ride roughshod through*, *run roughshod on*, *run roughshod of*, and *run roughshod in*. And there is one example of a passive use:

(108) BC was riddled by Miami backup quarterback Scott Covington, who [...] was run roughshod by running backs Dyral McMillan and freshman Edgerrin James, who [...]

sell like hotcakes

The idiom *sell like hotcakes*, which roughly means 'sell very fast', is canonical in 82% of its occurrences. Variations include *go like hotcakes*, *going out the door like hotcakes*, and *(Vanguard) has been selling index funds like hotcakes*.
Note that for speakers for whom *hotcakes* is an independent lexical entry this expression is a simile, so there is nothing idiomatic about this expression for such speakers.

get a word in edgewise

The idiom *get a word in edgewise* is canonical in 89% of its occurrences (17 out of 19). The two varied occurrences are *get a few words in edgewise* and *wedge a word in edgewise*.[23]

out of wedlock

The word *wedlock* occurs by itself 5% of the time: 468 out of 493 occurrences are of the form *out of wedlock*, but 8 examples are of the form *outside (of) wedlock* and 17 other examples contain *wedlock*.
However, at least a few speakers reject the corpus examples containing *wedlock* by itself like those in (109), and for those speakers *wedlock* is a word that occurs only as part of this idiom.

(109) a. Her latest flirtation with wedlock is on the rocks.
b. French couples shun wedlock.

91% of the occurrences of *wedlock* are about births, although only 35% are in the form of the collocation *born out of wedlock*. The other examples are of the form *(have) child(ren)/daughter(s)/son(s)/babies out of wedlock*, *birth(s) out of wedlock*, etc. Many of the remaining examples are about related things like *sex outside wedlock*. There is no variation internal to the idiom *out of wedlock* or internal to the collocation *born out of wedlock*.

[23]I also looked at the variant of this idiom involving *edgeways*. But it is too infrequent to make statistical observations about—there are only two occurrences of *edgeways* in the corpus. One of them is *I couldn't get a word in edgeways* and the other is *Dennis Rodman worked a couple of words in edgeways*.

3.5 A Random Sample of V+NP Idioms

In order to get reliable data about the average variability of idioms, it is necessary to look at a random sample of idioms. The Collins Cobuild Dictionary of Idioms marks the 750 idioms that are most frequent in the 'Bank of English'.[24] Many of these are just idiomatic noun phrases (e.g. *a track record*) or prepositional phrases (e.g. *in a vacuum*), and comparing their degree of variation with that of verb phrase idioms would not be meaningful because there are not as many possible ways in which such phrases can be varied. The same is true for verb phrase idioms when there is only an adjective involved, as in *see red* or *sit tight*.

The opposite is true for verb phrase idioms that involve variable complements (like *give someone free rein*) because these can potentially be varied in more different ways, such as dative alternation or topicalization of the variable element. More generally, any idiom involving more than one complement (e.g. *rub shoulders with someone*) has the potential for more variation, because the second complement can potentially be omitted or varied as well. For idioms involving prepositional complements (e.g. *get into gear*) there is the possibility of varying (e.g. *in* for *into*) or even omitting the preposition (*follow* (*in*) *someone's footsteps*), and if the prepositional phrase is not a complement, as in *read between the lines*, that also increases the possible types of variation.

So I excluded these types of idioms and selected a random sample[25] of 25 idioms from the resulting list of the 130 most frequent idioms that consisted only of a verb followed by a noun phrase. Below I have classified these randomly selected idioms

[24]This dictionary was prepared by British lexicographers and the Bank of English contains British English as well as American English text, so these frequencies may not be similar to those in the North American News Text Corpus. However, the dictionary includes information about the differences some idioms exhibit between British and American English. In the idioms from the random sample no such differences were noted, with the exception of the different spelling for *mo(u)ld*.

[25]I used a perl script that generates random numbers between 1 and 130.

into decomposable and non-decomposable idioms.[26] As discussed above, the boundary between these two categories seems to be fuzzy, so that not all speakers will agree with this classification. In particular, for some idioms for which a semantically decomposable interpretation is possible, some speakers may prefer a non-decomposable interpretation instead.

Decomposable idioms (16):

- *turn the tables* ('reverse the positions')
- *call the shots* ('make the decisions')
- *deliver the goods* ('deliver the expected')
- *lose face* ('lose status')
- *make waves* ('attract attention')
- *run the show* ('control the situation')
- *pay dividends* ('bring advantages')
- *sound the death knell* ('herald the end')
- *break the mold* ('escape the pattern')
- *lose ground* ('lose value', 'lose the advantage')
- *strike a chord* ('touch sensibilities')
- *rear its (ugly) head* ('manifest its (negative) presence')
- *break the ice* ('end the silence')
- *level the playing field* ('equalize the situation')
- *lead the field* ('be ahead of the competition')
- *take a back seat* ('take a subsidiary position')

Non-Decomposable idioms (6):

- *hit home* ('affect personally')
- *speak volumes* ('reveal a lot')
- *close ranks* ('be supportive')
- *look the other way* ('deliberately ignore (it)')

[26] This classification was done by thinking about possible paraphrases, before conducting the corpus study. However, at that time I only took into account whether or not a V+NP paraphrase existed, not whether it matched the metaphor. Therefore I initially misclassified the idioms *clear the air* and *bite the bullet*.

- *clear the air* ('address the misgivings')
- *bite the bullet* ('accept the situation')

Many of these expressions have not been discussed in the linguistic literature before, and as we will see below it is not clear whether all of them are really idioms, or whether some, e.g. *pay dividends*, may instead be collocations. A separate analysis of results excluding these expressions will be given. But in order to avoid affecting the randomness of the sample as much as possible I studied all of these expressions, appealing to the authority of the COBUILD Dictionary of Idioms, which delimits the scope of the dictionary as follows:

> An idiom is a special kind of phrase. It is a group of words which have a different meaning when used together from the one it would have if the meaning of each word were taken individually. [...] Idioms are typically metaphorical: they are effectively metaphors which have become 'fixed' or 'fossilized'. [...] The COBUILD Dictionary of Idioms [...] includes traditional English idioms such as *spill the beans* and *a red herring*. It also includes a number of expressions which can be considered 'semi-idioms': some very common multi-word metaphors such as *the acid test* and *brownie points*; metaphorical proverbs such as *every cloud has a silver lining* and *in for a penny, in for a pound*; common similes such as *white as a sheet* and *old as the hills*; and some other expressions which have a strong pragmatic meaning, such as *famous last words* and *that's the way the cookie crumbles*. We have deliberately avoided including other kinds of fixed expression such as *in fact* and *at least*, or greetings and other fixed formulae such as *how do you do* and *excuse me*.

There were three further idioms in the random sample which I did not analyze for various reasons. The first of these idioms was *fly the flag* ('show pride in the group'). I searched for this idiom in the corpus and found only eight canonical occurrences that are clearly idiomatic, like:

(110) [...] what's the point of flying the flag for Shakespeare.

The other 66 'canonical' occurrences of this expression involved literal flags, although some of these are clearly about expressing pride as well. But for many examples, such as the one in (111), it is not clear whether the writer intended the idiomatic meaning.

(111) I almost forgot to fly the flag for the Fourth of July.

Therefore I decided it was not worth sifting through the remaining 1273 matches of the corpus query, which are mostly irrelevant (e.g. *a flag flies*). Most probably I would not be able to tell with a high enough degree of certainty which ones are idiomatic, or find enough clear instances of this idiom to make the data statistically meaningful.

The second idiom I excluded was *score points* ('gain an advantage'). This query produced the largest number of matches (12279) of any of the idiom queries, and apart from the fact that it could take days to sift through all these data, it would be really hard to separate some of the more literal sports uses from the idiomatic ones, because scoring points in a game also gives one an advantage, so in some cases the idiomatic meaning may be intended:

(112) What Armstrong noticed immediately about coming from the Bears "is that when this team gets a turnover, it can score points off of them."

This example also shows that I would not be able to automatically delete all sentences containing words like *game* and *team*. Furthermore, I did not realize when I included this idiom in the list of 130 V+NP idioms that the preposition *off* was part of the idiom with this meaning, *score points with* being a separate idiom with a different meaning ('gain favor with').[27] There are also some corpus examples of this idiom without the preposition, or with a different preposition like *over*, but to the extent that the preposition is part of the idiom it should have been eliminated using my criteria for making the list of 130 V+NP idioms. Note also that in many examples that do not involve a preposition, it is not clear without a larger context whether the 'gain an advantage' meaning or the 'gain favor' meaning is intended:

(113) Dole accused Clinton of using the nomination to score political points on abortion.

[27]These two idioms are listed separately in the Collins dictionary, and it is *score points off* which has the special marker for high frequency idioms.

This could be an example of 'scoring points off the politicians who are against the nominee' and of 'scoring points with the people who share Clinton's view on abortion' at the same time.

The final idiom I excluded is *play ball* ('cooperate'). Again there would be a large number of matches to sift through (3724), and again it is not clear that the idiomatic meaning can always be identified without looking at a larger context:

(114) Because jocks play ball, we play ball with them, and blindly make them our icons.

And again, if the preposition *with* is part of the canonical form of this idiom, it should have been excluded from the list of the 130 V+NP idioms anyway.

In the next two sections I describe the results of the corpus study of the 22 randomly selected idioms.

3.5.1 Decomposable Idioms

turn the tables

69% of the occurrences of this idiom, which roughly means 'reverse the positions', are in their canonical form: 357 out of 518 show no variation other than inflection of *turn*.

There are 161 non-canonical examples of this idiom, which represents 31% of the occurrences in the corpus.

- In 1 example *the table* is singular
- In 10 examples there is no specifier (*turn tables*)
- 3 examples have a modifier (*rhetorical*, *fiscal*, *political*)
- 1 example involves a compound noun (*the filibuster tables*)

65 examples are passive:

- 15 examples are of the form *the* (modifier) *tables* (auxiliary) (adverb) *have been turned* (e.g. *the moral tables have been turned*)
- 2 examples are of the form *the tables would/could be turned*

- 48 examples are of the form *the* (modifier) *tables are/were* (adverb) (*being*) *turned* (e.g. *the tables were promptly turned*, *now the partisan tables are being turned*, *the political tables are turned this year*, *if the tables were turned*) (some of these are adjectival passives—Wasow (1977))

65 examples are inchoative:[28]

- 23 examples are of the form *the* (modifier) *tables* (adverb) *turned* (e.g. *the tables suddenly turned*)
- 34 examples are of the form *the* (modifier) *tables* (adverb) (auxiliary) *have* (adverb) *turned* (e.g. *the technical tables may have turned*, *the tables have now turned*)
- 8 examples are of the form *the* (modifier) *tables* (*will*) (adverb) *turn* (e.g. *the tables soon turn*)

3 examples involve raising (one is also inchoative and one is also passive):

- the tables appeared to have turned
- the tables appeared to turn
- the tables seem utterly turned

13 examples are unusual in other ways:

- the tables will keep turning
- the tables (modifier) get turned
- Many art-world experts attribute this to the tables' having turned [. . .]
- whose tables turn
- when did the tables turn?
- turning of the tables (5)
- turning of the corporate tables
- turn of the tables
- the drop-leaf table, however, is turning

[28]Note that for semantic reasons inchoative uses are not possible for all idioms.

call the shots

79% of the occurrences of this idiom, which roughly means 'make the decisions', are in their canonical form: 467 out of 589 show no variation other than inflection of *call*. There are 122 non-canonical examples of this idiom, which represents 21% of the occurrences in the corpus.

- In 9 examples *shot* is singular (*the shot* (6), *that shot*, *his shot*, *my own shot*)
- In 4 examples there is no specifier (*call shots*)
- In 83 examples there is a different specifier (*all the* (30), *most of the* (10), *his own* (10), *their own* (8), *many of the* (5), *her own* (3), *some of the* (2), *its own* (2), *your* (2), *their* (2), *our*, *his*, *plenty of the*, *all of the*, *a lot of*, *more of the*, *fewer and fewer*, *virtually all the*, *NATO's*)
- 11 examples have a modifier (*major* (3), *important*, *final*, *artistic*, *daily*, *Democratic*, *right*, *creative*, *wrong political*)
- 6 examples involve a compound noun (*space shots*, *basketball shots*, *IRA shots*, *camera shots*, *public policy shots* (e.g. *a female is calling the space shots 400 years in the future*)
- 4 examples involve both a different specifier and a modifier (*all the political shots*, *most of the political shots*, and *all the linguistic shots*)

5 examples are unusual in other ways, including 4 passives:

- a lot of the shots are being called by the Bosnian Serbs right now
- Basically, a lot of shots are being called by the Bosnian Serbs
- the shots clearly will be called from San Francisco
- the shots clearly are being called from GM headquarters
- It's his shot to call

deliver the goods

84% of the occurrences of this idiom, which roughly means 'deliver the expected' or 'deliver the necessary', are in their canonical form: 148 out of 176 show no variation other than inflection of *deliver*.

There are 28 non-canonical examples of this idiom, which represents 16% of the occurrences in the corpus.

- In 2 examples there is no specifier (*deliver goods*)
- In 6 examples there is a different specifier (*the same*, *the whole bill of*, *such*, *their*, *our*, *the best of his*)
- 12 examples have a modifier (*scenic*, *emotional*, *creative*, *economic*, *social*, *good enough*, *political*, *diabolical*, *promised*, *public*, *entertaining*, *spectacular*)
- 3 examples have a both specifier and a modifier (*the film's valuable goods*, *some delicious goods*, *several kinds of reliable goods*)
- 2 examples involve a compound noun (*the action goods*, *the thriller goods*)

4 examples are unusual in other ways:

- we really haven't seen the goods delivered
- if the goods aren't delivered
- crass exploitation as the delivered goods

Note that for this expression the line between idiom and non-idiom is a bit blurry when the subject is a person, i.e. able to 'deliver' things like services and speeches, as in (115).

(115) It intensifies the pressure on Dole to deliver the goods in the Senate.

It is not always clear whether the idiomatic meaning is intended in such examples, so I am not totally sure I classified each corpus example correctly. More specifically, I may have counted some literal uses as idiomatic, as I counted examples involving non-physical 'goods' like (115) as idiomatic.

lose face

85% of the occurrences of this idiom, which roughly means 'lose status', are in their canonical form: 116 out of 137 show no variation other than inflection of *lose*.
There are 21 non-canonical examples of this idiom, which represents 15% of the occurrences in the corpus.

- In 10 examples there is a specifier (*too much* (4), *more* (2) *a lot of*, *some*, *any*, *all*)

- 5 examples have a modifier (*political*, *great*, *social*)

3 examples involve coordination with non-idiomatic items:[29]

- And everywhere it is losing pride and face
- The longer they go without a collective-bargaining agreement, the more money and face both sides will lose.
- The infamous catamaran debacle of 1988 saw the New Zealanders lose both face and 0-2 on the water.

3 examples are unusual in other ways, including one occurrence across a relative clause boundary:

- [...] the United States is trying to regain face lost when President Clinton backed down last year [...]
- has far less face to lose
- [...] you have the problem of face—and China hates to lose any

make waves

77% of the occurrences of this idiom, which roughly means 'attract attention', are in their canonical form: 187 out of 243 show no variation other than inflection of *make*. There are 56 non-canonical examples of this idiom, which represents 23% of the occurrences in the corpus.

- In 27 examples there is a specifier (*any* (6), *more* (5), *some* (3), *a lot of* (2), *the most* (2), *no*, *such*, *lots of*, *more than a few*, *too many*, *far more*, *plenty of*, *as many*, *many*)
- 19 examples have a modifier (*big* (6), *political* (2), *huge*, *considerable*, *giant*, *scientific*, *additional*, *new*, *maximum political*, *similar*, *real*, *even greater*, *even bigger*)
- 3 examples have both a specifier and a modifier (*its biggest*, *some early*, *such big*)
- 2 examples involve a compound noun (*sales waves* and *marketing waves*)

[29]These examples suggest that only *face* is an idiomatic word and *lose* has its literal meaning, confirming a paraphrase like 'lose status'.

5 examples are unusual in other ways, including one coordination with non-idiomatic material, 1 passive, and 2 occurrences across relative clause boundaries:

- [...] the stage is set for Capt. Paula Meara, a 22-year veteran of the force, to make history, and then some waves.
- Don't expect this show to make many, uh, waves as far as ratings are concerned.
- [...] her activism in a slot where few waves have been made since the late 1970s.
- [...] the waves they made.
- The waves Japanese authorities are making in the currency markets [...]

run the show

78% of the occurrences of this idiom, which roughly means 'make the decisions', are in their canonical form: 301 out of 386 show no variation other than inflection of *run*. There are 85 non-canonical examples of this idiom, which represents 22% of the occurrences in the corpus.

- In 8 examples there is a different specifier (*this* (3), *that*, *his*, *its*, *their*, *much of the*)
- 19 examples have a modifier (*whole* (11), *entire* (3), *political*, *corporate*, *authoritarian*, *international monetary*, *civil service legislative*)
- 6 examples involve a compound noun (*Manhattan Beach show*, *Senate show*, *vulture-fund show*, *CFE show*, *GOP show*, *entire sports show*)
- 2 examples involve both a different specifier and a modifier (*this whole*, *a sloppy*)

49 examples are of the form *run one's own show*.

- Analysts think the company's ability to run its own show will improve its performance.
- Before you try it, though, ask yourself if you're suited to run your own show away from the social interaction and supervision of an office.
- For the doctors, however, running their own show could represent quite a change, say health specialists.

1 example is passive.

- This show is not run by human reckonings, senior management and things like that.[30]

Note that I may have failed to eliminate some occurrences of this expression that are about literal *shows*, because sometimes it is not possible to tell without looking at a broader context whether the idiomatic meaning is intended. Even with modified examples like *run the entire sports show* that sound literal at first glance one sometimes has to look at the larger context:

(116) Charley Pell and Norm Sloan may have missed Monday's announcement on Debbie Yow, 42, being named Maryland's new athletic director. Of the 107 Division I-A schools that play major college football, there are only three other women **running the entire sports show**.

I did however eliminate obvious occurrences of *running the TV show* etc. If I overlooked literal uses I am a bit more likely to have done so for the canonical forms, so the actually percentage of variability for this idiom may be a little higher than 22%.

pay dividends

52% of the occurrences of this idiom, which roughly means 'bring advantages', are in their canonical form: 218 out of 418 show no variation other than inflection of *pay*. There are 200 non-canonical examples of this idiom, which represents 48% of the occurrences in the corpus.

- In 15 examples *dividend* is singular, and all but two involve a specifier and/or modifier: (*a*, *much of a*, *one*, *first* (2), *its first*, *his first*, *the unexpected*, *a valuable*, *its principal political*, *a big fiscal*, *even a single ratings*, *the highest*)
- In 21 examples there is a different specifier (*some* (8), *no* (3), *any* (2), *its own* (2), *few*, *its*, *a lot of*, *the kind of*, *the most*, *much more*)
- 141 examples have a modifier (*big* (31), *huge* (17), *immediate* (12), *political* (9), *handsome* (8), *rich* (5), *quick* (4), *enormous* (3), *big political* (3), *economic and strategic* (3), *great* (2), *other* (2), *major* (2), *equal* (2), *large* (2), *higher* (2), *good* (2), *important* (2), *big economic* (2), *economic*, *double*, *hefty*, *early*,

[30]This example is about Mother Teresa's work in Calcutta.

later, *healthy*, *greater*, *golden*, *social*, *discernible*, *high*, *substantial*, *fat*, *solid*, *early*, *instant*, *staggering*, *untold economic*, *big fiscal*, *large fiscal*, *great academic*, *heavy political*, *handsome political*, *considerable political*, *large political*, *domestic political*, *economic and political*, *large economic and political*)

- 11 examples have a different specifier and a modifier (*some political* (3), *the highest* (2), *some big*, *some fast*, *such rich*, *such huge*, *the highest of*, *its greatest*)
- 5 examples involve compound nouns (*World Cup dividends*, *substantial public-health dividends*, *year-round productivity dividends*, *bid dividends*, *real-world dividends*)

7 examples are unusual in other ways, including 2 occurrences across relative clause boundaries and 3 with intervening datives (*pay someone dividends*).

- He's the ultimate striver—and what dividends his strivings have paid.
- which nonetheless paid what Brown and her advisers believe are important dividends.
- The singers and their pastor have no doubts about the dividends their presence has paid for the families of the victims.
- [...] the revenue-sharing plan for the WAC pales in comparison to recent dividends paid by SWC membership.
- [...] the fact that there is a strong Zenith presence will pay us dividends in the future.
- [...] training which paid the city an obvious peace dividend at the end of the second Rodney King trial.
- Mr. Clinton's determined effort [...] paid him critical dividends at the polls.

It is not totally clear whether this expression may be a collocation rather than an idiom. The word *dividend* can be used metaphorically with the same meaning ('advantage', 'return') with other verbs, as in (117):

(117) a. [...] Kemp's self-appointed status as a racial and religious peacemaker could **provide political dividends** [...]

b. [...] Republicans could **reap political dividends** [...]

c. [. . .] its few solid achievements have **brought scant political dividends**.

d. Rabin **hopes for political dividends** as well.

And the Collins COBUILD dictionary lists *dividend* as having the meaning 'an extra benefit that you did not expect to get'. However, it is not clear whether it can have this meaning in general: (?*lose dividends*). And neither **pay advantages* nor ?*pay returns* sound particularly plausible, so it is not clear that *pay* can have exactly the same meaning in other constructions. But *pay off* seems to be closely related, and can in fact be substituted with little change in meaning for *pay dividends* in typical examples like those in (118):

(118) a. Careful plant selection will pay dividends in such conditions.

b. Kohl's last visit already has paid dividends.

c. His strategy appears to be paying political dividends.

d. [. . .] taking time to listen is sure to pay handsome dividends.

It is clear that this idiom is very transparent and that the underlying metaphor is both active and common. It is also likely that it may be collocation for some speakers. So it is not surprising that the degree of variation of this expression is not quite like that of other idioms.

sound the death knell

66% of the occurrences of this idiom, which roughly means 'herald the end', are in their canonical form: 73 out of 110 show no variation other than inflection of *sound*. There are 37 non-canonical examples of this idiom, which represents 34% of the occurrences in the corpus.

- In 5 examples there is no specifier (*sound death knell*)
- In 16 examples there is a different specifier (*a* (12), *its* (2), *the organization's*, *the B-2's*)
- 1 example has a modifier (*final*)

3 examples are passive:

- A death knell for California Gov. Pete Wilson's presidential ambition was sounded at an event that received little coverage.

- The death knell has been sounded many times for newspapers.
- National Hockey League fans across Canada fear the death knell has been sounded for the Quebec City Nordiques [...]

One example goes across a relative clause boundary:

- [...] defy the death knell much of Wall Street was sounding a year ago.

8 examples are inchoative, so *the death knell* is the subject, and in one of these examples *knells* is plural:

- Is the death knell finally sounding for the "tax on honesty" [...]
- [...] the death knell sounded for the diesel-powered American car at the end of the 1985 model year.
- For now, the death knell is sounding on the U.S. Festival [...]
- By the late 1950s, the death knell was sounding for what once looked like a new and powerful art form.
- The death knell will sound for Prime Minister Silvio Berlusconi's government Wednesday [...]
- The death knell for "Playhouse 90" and all similar enterprises sounded when Robinson was replaced by James Aubrey [...]
- The death knell, he says, began to sound for Crowded House after "Together Alone," the band's fourth album, was released [...]
- [...] because of sharply falling sales, death knells have been sounding for the classical record industry.

3 examples are unusual in other ways:

- "This may be the sounding of the death knell for this system."
- [...] traders said the sounding of the death knell may have been premature.
- Bossi sounded what seemed to be the death knell for Berlusconi's seven-month-old government [...]

Note that *death knell* can occur with the same meaning ('end') in other constructions (e.g. *be the death knell*)—in fact there are 191 such examples in the same corpus.[31]

[31]The word *knell* occurs 97% of the time as *death knell*.

However, *sound* cannot occur with this meaning ('herald') outside the idiom (??*sound the end*, although *sound the warning/alarm/retreat* seems to be a related use), so this expression is still an idiom according to my definition. It clearly matches a more intuitive definition of idiom because both parts are figurative, and are established in this combination. Note also that in many examples the meaning seems to be closer to *cause the end* or at least *contribute to the end*:

(119) a. Oakman said he did not think that annotated literature on CD-ROM would sound the death knell for books.

b. I think the decision sounds the death knell for affirmative action and minority set-asides [...]

c. Some politicians said Clinton's reference in Bonn Monday to a "unique" relationship between Germany and the United States sounded the death knell of the "special relationship" between Washington and London.

d. [...] it seemed reasonable for Mr. Clinton's advisers to assume the study would sound the B-2's death knell.

The idiom may have yet another meaning, as **the end heralds* cannot be the meaning of the inchoative uses. Perhaps this just indicates that *herald the end* is not a very good paraphrase, but I am not sure whether there is a better one.

break the mold

61% of the occurrences of this idiom, which roughly means 'escape the pattern', are in their canonical form: 102 out of 168 show no variation other than inflection of *break*.[32]

There are 66 non-canonical examples of this idiom, which represents 39% of the occurrences in the corpus.

- In 11 examples *molds* is plural and all except two occurrences of *break molds* also have specifiers and/or modifiers (*some molds* (2), *the molds*, *any molds*, *old molds* (2), *all previous molds*, *political molds*, *the traditional molds*)

[32]I counted the 16 occurrences of *mould* with its British spelling the same way they would have been counted in the American spelling, i.e. some of them were counted as canonical and others as non-canonical.

- In 11 examples there is a different specifier (*that* (6), *a* (2), *his*, *this*, *your own*)
- 15 examples have a modifier (*conventional*, *usual*, *current*, *British*, *boxy*, *communist*, *political*, *corrupt*, *royal*, *old*, *consensual*, *long-winded*, *old political*, *vice presidential*, *male-dominated political*)
- 1 example has a different specifier and a modifier (*the show's traditional*)
- 19 examples involve a compound noun (*the family mold*, *the immigrant mold*, *the slapstick mold*, *the slacker mold*, *the ad mold*, *the Hollywood mold*, *the Wills mold*, *the Astor mold*, *the "SNL" mold*, *the post-war mold*, *the suit-and-tie mold*, *the campaign-coverage mold*, *the action-comedy mold*, *the gangsta-flick mold*, *the post-war mould*, *the usual Hollywood mold*, *the cramped Hollywood mold*, *the traditional John Thompson mold*, *'the pretty-face, perfect-hair, former-jock mold'*)

9 examples are unusual in other ways, including one example where *mold* is the subject, four passives, two occurrences across relative clause boundaries, and one pronominal reference:

- a breaking of the mold
- That mold is going to break again during this trial [. . .]
- [. . .] the mould has been broken. (2)
- If the mold could be broken, it would be helpful for everybody [. . .]
- The mold of these clothes—be it the delicate metallic lace dress, the rabbit fur coat, the fur-trimmed trench or the black watch plaid suit—wasn't broken by Jacobs.
- [. . .] that's precisely the mold "Prime Suspect" broke to become great.
- [. . .] a mold that badly needs breaking.
- It's up to us who don't fit the Generation X mold to break it [. . .]

Note that *mold* can have the same meaning in *fit the mold*. But this is another conventionalized idiom which fits the same conventionalized metaphor, and *mold* cannot mean *pattern*, *prototype* or *stereotype* in other places (**follow the mold*, **match*

the mold, **study/observe/notice a mold*).[33] So this expression clearly has to be represented at the phrasal level. But it is probably also necessary to express the higher-level metaphor, as it is quite plausible that other verbs with related meanings (such as *shatter*, *crack*, *break out of*) can be interpreted without problems. This might explain why this idiom is somewhat more variable.

lose ground

About 70% of the occurrences of this idiom, which roughly means 'lose value' or 'lose the advantage', are in their canonical form: about 1653 out of 2350 show no variation other than inflection of *lose*.[34]
There are 697 non-canonical examples of this idiom, which represents 30% of the occurrences in the corpus.

- 189 examples are the NP *lost ground*
- In 220 examples there is a different specifier (*some* (65), *more* (35), *a little* (18), *a lot of* (15), *much* (14), *the most* (11), *further* (9), *so much* (7), *any* (6), *even more* (5), *no* (3), *a great deal of* (2), *a bit of* (2), *enough* (2), *too much* (2), *as much* (2), *some of the* (2), *a*, *the*, *his*, *little*, *less*, *lots of*, *all the*, *all of the*, *a fair amount of*, *an inch of*, *significant amounts of*, *quite a bit of*, *some of that*, *most of his*, *much of their*, *all that*, *how much*, *any more*, *some more no further*)
- 46 examples have a modifier (*major* (5), *considerable* (5), *substantial* (4), *economic* (4), *political* (4), *significant* (3), *valuable* (2), *strategic* (2), *moderate* (2), *electoral*, *immunological*, *real*, *financial*, *conservative*, *legislative*, *new*, *important*, *popular*, *modest*, *hard-won*, *only modest*, *political and economic*, *economic and political*, *precious, hard-won*)

[33]Note that while this metaphor may seem very transparent to native speakers of English, and is probably not too hard to figure out for non-native speakers, the way it is expressed is clearly conventionalized. It is completely impossible to use a similar expression in German: (**Er* (*zer*)*bricht die übliche Gussform/Form*)

[34]The reason I am not as sure about the numbers with this idiom is that I did not look at all the 1842 'canonical' matches manually. Some of these are the NP *lost ground*, and I only estimated those by grepping for *the lost ground*, *some lost ground*, *any lost ground*, *up lost ground*, *back lost ground*, *regain/s/ed/ing lost ground*. I also did not manually delete all literal army uses.

- 11 examples have a specifier and a modifier (*the precious* (5), *the athletic*, *any artistic*, *some political*, *so much political*, *a great deal of political*, *a lot of f------*)

220 examples go across clause boundaries. A few of them are given below:

- Treasury bonds recovered some ground lost earlier today [. . .]
- [. . .] the dollar recovered all the ground lost against the yen [. . .]
- The U.S. stock market yesterday recovered all the ground it lost on Friday [. . .]
- [. . .] Republicans have regained ground they lost [. . .]
- Sales of new products more than made up the ground lost by Tagamet [. . .]

There are 2 examples of coordination:

- What ground she breaks, or loses, may very well be followed by the others.
- Tom Lasorda was asked to consider these six wins in nine games, in places where the Los Angeles Dodgers typically lose, both games and ground.

9 examples are unusual in other ways:

- All that ground is lost.
- But Shalala noted that ground is being lost, for example, in the battle against obesity [. . .]
- [. . .] hard-gained ground in battling narcotics may be lost.
- [. . .] which ones are gaining ground and which are losing.
- Several of the chip-equipment makers that gained ground Friday lost a little in this session.
- [. . .] I'm not convinced we're gaining ground and we may well be losing a bit.
- Such acts of violence, he said, had lost the Catholics "the morally advantageous ground" they had won when the Protestants engaged in the provocative marches.
- It will also, in 1996, become clear to the torchbearers of Deng's reforms how much ground has been lost to conservatives in the previous two years.
- Scarcely believing how much ground they have lost in the past six months, [. . .]

Note that the unusually high occurrence of this idiom across clause boundaries is partially explained by the fact that *ground* does not have to be 'licensed' by *lose*, because words like *(re)gain*, *recover*, and *make up* also conventionally occur with *ground*, and fit the metaphor.[35] But *ground* cannot mean *value* or *advantage* outside this group of metaphorical idioms (**the dollar has a lot of ground this year*, **the dollar has* (*a*/*the*) *ground compared to the Euro*, **the ground of the dollar is increasing*).

strike a chord

52% of the occurrences of this idiom, which roughly means 'touch sensibilities', are in their canonical form: 395 out of 754 show no variation other than inflection of *strike*. There are 359 non-canonical examples of this idiom, which represents 48% of the occurrences in the corpus.

- In 1 example there is no specifier (*strike chord*)
- In 11 examples *chords* is plural
- In 45 examples *chords* is plural and there is an added specifier or modifier (*some* (2), *at least some*, *responsive* (3), *deep* (3), *deeper* (2), *strong*, *familiar*, *regional*, *single*, *eerie*, *resonant*, *universal*, *particular*, *emotional*, *mainstream*, *winning*, *promising*, *some deep*, *some deep emotional*, *some similarly eerie*, *some other sensitive*, *some angry*, *some pleasing*, *some resonant*, *such evocative*, *such vibrant*, *the same responsive*, *the most somber*, *the same magical*, *few remembered*, *a few resonant*, *all the right*, *many popular*, *many traditional bipartisan*, *many of the cultural and family-values*, *resonant emotional*, *powerful historical*, *'huge, clanging'*, *heart*)
- In 21 examples there is a different specifier (*such a* (10), *this* (2), *that*, *no*, *many a*, *not only a*, *some sort of a*, *less of a*, *little*, *as emotional a*, *as responsive a*)
- 238 examples have a modifier (*responsive* (49), *deep* (12), *emotional* (10), *sympathetic* (9), *popular* (7), *resonant* (6), *powerful* (5), *familiar* (5), *positive*

[35]This metaphor is probably understandable to anyone from a society in which people fight over the ownership of land. But it there are still different ways in which it could have been conventionalized. For example, there is a similar conventionalized metaphor in German, but it uses a different expression involving a partitive (*der Dollar hat* ***an*** *Boden verloren*, **der Dollar hat Boden verloren*).

(5), *particular* (4), *similar* (4), *special* (4), *particularly sensitive* (3), *patriotic* (3), *deep emotional* (2), *receptive* (2), *common* (2), *deeper* (2), *new* (2), *dark* (2), *very strong* (2), *populist*, *forgiving*, *appreciative*, *provocative*, *encouraging*, *massive*, *negative*, *resonating*, *real*, *idealistic*, *defiant*, *feminist*, *harmonious*, *greater*, *recognizable*, *nationalist*, *national*, *tender*, *ironic*, *strong*, *warm*, *poignant*, *American*, *universal*, *major*, *skeptical*, *resounding*, *appropriate*, *annoying*, *sensitive*, *different*, *centrist*, *unsympathetic*, *electoral*, *stronger*, *nostalgic*, *compelling*, *mellow*, *divisive*, *poetic*, *uncomfortable*, *hollow*, *mournful*, *pleasant*, *momentous*, *political*, *valuable*, *sour*, *dissonant*, *philosophical*, *harmonic*, *phenomenal*, *racial*, *visceral*, *timely*, *much-needed*, *more poignant*, *very different*, *very powerful*, *very sensitive*, *ready public*, *even richer*, *important political*, *different psychological*, *strong emotional*, *strong political*, *strong nationalistic*, *deeply resonant*, *deeper emotional*, *particularly emotional*, *particularly dissonant*, *particularly appealing*, *particularly ominous*, *particularly responsive*, *surprisingly responsive*, *especially nervous*, *clearly contemporary*, *shockingly poignant*, *distinctly conservative*, *especially harsh*, *broadly responsive*, *universal romantic*, *loud clear*, *clear, vibrant*, *deep, resonant*, *powerful, favorable*, *bittersweet, oddly conflicted*, *deep and familiar*, *responsive or responsible*, *more widely felt*, *'very, very deep'*)

- 24 examples have a different specifier and a modifier (*the right* (3), *such a responsive* (2), *the resonant*, *the same*, *the wrong*, *the loudest*, *the strongest*, *the deepest*, *the right cultural*, *the first hopeful*, *the only familiar*, *the same nesting*, *the proverbial responsive*, *just the right*, *such a deep*, *such a harmonious*, *such a resounding*, *much of a responsive*, *some kind of responsive*, *some emotional*, *some sort of responsive and admiring*)
- 4 examples involve compound nouns (*a death chord*, *the death chord*, *a Third-World chord*, *the same rocker chord*)

11 examples occur across relative clause boundaries, and one of them is also passive:

- The ethnic chords Tonelli strikes are indeed vestigial [...]
- [...] the popular chord that Buchanan has struck [...]
- [...] the primeval chord that giraffes strike in men's souls.

- [...] the chord it strikes among misfits around the globe [...]
- [...] the chord Mrs. Clinton strikes with ethnic voters [...]
- [...] the chord it strikes among the unhappily wed [...]
- [...] the uncompromising chord he struck when his party first took control of Congress [...]
- [...] the moral chord he has struck [...]
- [...] is another chord marketers are striking.
- [...] the responsive chords the Republicans struck in the 1994 congressional elections.
- [...] the "responsive chord" struck by the Million Man March [...]

4 examples are unusual in other ways, including one passive:

- the chord of continuity
- the chord that framed the story
- no matter what chord it strikes in a user or listener
- The real emotional chord, though, was struck with Yad Vashem

Note that 291 (39%) of the occurrences of this expression are followed by *with*. Some typical examples are given in (120).

(120) a. Mr. Erbakan's nationalism strikes a chord with many Turks, [...]

b. They may give to charities that strike a chord with them, because of family illnesses.

It is quite striking how much more variable than average this idiom is, especially considering that its meaning could just as well be non-decomposable in the unmodified examples, given that *resonate with* is the best paraphrase for many of these examples. Most of the adjectives can be given an external modification interpretation like *resonate emotionally*, and it is hard to tell these apart from internal modification readings like *touch emotional sensibilities* or *trigger an emotional response*. This shows that this metaphor, although conventionalized in this idiom, is not fossilized, and actually quite active—even the paraphrase *resonate* follows the same metaphor, and some other related verbs like *touch* can be used instead of *strike* in the idiom.

Nevertheless, the metaphor *strike a chord* is clearly conventionalized. It fits my definition of idiom because *chord*(*s*) by itself cannot be used to mean *sensibilities* (#*be aware of her chord*(*s*)) or *response* (#*cause/trigger a chord*). Whether or not it is called 'idiom', it needs to be represented at the phrasal level.

However, it is also clear that this idiom is treated as decomposable by many speakers. This is suggested by the high rate of modification of *chord*, and some clear cases are given in (121). In these examples it appears that *strike a racial chord* has to mean *touch racial sensibilities* rather than to *resonate in a racial way*, and *strike a Third-World chord* most plausibly means *touch Third-World sensibilities*.

(121) a. Much of what he says on subjects like crime and welfare has an undertone that strikes a racial chord among African Americans.

b. Samper struck a Third-World chord and received a sustained ovation in the 185-member chamber.

For examples that are modified with adjectives that are more metaphorical, like those in (122), *resonate* is a better paraphrase than *touch sensibilities* (*resonate in a harmonious way*, *resonate in a dissonant way*).

(122) a. The world leaders were unlikely to strike such a harmonious chord in Moscow Tuesday

b. But in Utah, the issue strikes a particularly dissonant chord.

Note that there are other reasons to think that speakers differ in how they analyze this idiom. For the person who used *strike little chord*, idiomatic *chord* is obviously a mass noun, while it clearly is not used that way by most people. My two idioms dictionaries give quite different definitions ('make someone respond in an emotional way' vs. 'cause someone to remember something' or 'be familiar'). And when asking various native speakers I got a long list of different responses ('find something true in one's own experience', 'touch someone', 'be in harmony', 'resonate with').

For the adjective *responsive*, *resonate* is again the most plausible paraphrase: ?*resonate responsively*, **touch responsive sensibilities*, ??*trigger responsive responses*. However, this particular adjective is very frequent (it is included in 10% of the occurrences of the idiom) and is probably a remnant of the history of this idiom. In

fact, varied occurrences of this idiom like *strike the same responsive chord* and *strike some kind of responsive chord*, and in particular the two occurrences across clause boundaries suggest that this expression is decomposable after all, and that *strike a responsive chord* simply means *touch sensibilities*, without much of a contribution by *responsive*, except perhaps an intensifying one.

Note that when looking at *strike a responsive chord* as a conventionalized expression in its own right, it occurs in its canonical form 49 out of 66 times, i.e. 74% of the time. Of the remaining 688 occurrences of *strike a chord* 395 are canonical, i.e. 57%. This seems to be the right way of looking at the data, because the idiom *strike a chord* cannot be held accountable for the fact that such a closely related variant exists when considering its rate of variability.[36]

rear its head

74% of the occurrences of this idiom, which roughly means 'manifest its (negative) presence', are in their canonical form: 95 out of 128 show no variation other than inflection of *rear*. Note that I excluded all occurrences of *rear its ugly head* from the analysis, as it is clearly a conventionalized variant of the idiom rather than productive variation.[37]

Note that in nearly all of this idiom's occurrences the manifested presence is negative, even in the absence of *ugly*. Some typical canonical examples of the idiom *rear its head* are:

- Not everyone is so sure that inflation won't soon rear its head.
- There is a general uneasiness that rears its head before each Amgen report [...]
- Deadly mayhem rears its head at regular intervals.
- When racial bigotry rears its head [...]

[36]Consider what would happen if it became possible to say *come circle* instead of *come full circle*. Until such time as *full* is no more frequent than any other modifier (like perhaps *complete*), it would be meaningless to consider examples of people still using *come full circle* as 'varied'.

[37]Both the NTC and the Collins idioms dictionary list this variant. Given that it is conventionalized, it would not be meaningful to count the 67 occurrences of this idiom that include *ugly* as varied. Note that 53 of the occurrences of *rear its ugly head* (=79%) are canonical, i.e. they show no variation other than inflection of *rear*.

There are 33 non-canonical examples of this idiom, which represents 26% of the occurrences in the corpus.

- 12 examples involve *their*: 1 example is *their head*, 7 examples are *their heads*, and in 4 examples there is an additional modifier or compound (*their all-too-public heads*, *their three-dimensional heads*, *their scary, mind-bending heads*, *their gargoyle heads*)
- 18 examples have a modifier (*shaggy* (2), *wrathful*, *Gorgonic*, *gray*, *shiny*, *invisible*, *poisonous*, *relentless*, *pedantic*, *lovely*, *plastic*, *despicable*, *English*, *bandaged*, *non-conclusive*, *well-coiffed*, *grotesque little*)
- 3 examples involve a compound noun (*its Medusa head*, *its little Dynel head*, *its accountant's head*)

Note that several modifiers like *well-coiffed*, *gray*, *bandaged*, and *shaggy* are inspired by the literal meaning of *head*. But it is clear that the meaning is still idiomatic:

(123) a. Youth is rearing its shaggy head again on runways in Paris [...]

b. As we head into the "fin de siecle," extravagance – rendered politically incorrect by the recession – is rearing its well-coiffed head at nearly every show.

c. And when the International Olympic Committee rears its gray head [...]

d. The health issue reared its bandaged head [...]

Most of these examples work reasonably well with the idiomatic meaning (*shaggy presence*, *well-coiffed presence*, *gray presence*, *bandaged presence*). But they still need to be understood metaphorically, and it is likely that not just their production but also their comprehension involves the idiomatic metaphor.

break the ice

79% of the occurrences of this idiom, which roughly means 'end the silence/tension', are in their canonical form: 145 out of 183 show no variation other than inflection of *break*.

There are 38 non-canonical examples of this idiom, which represents 21% of the occurrences in the corpus.

- In 3 examples there is no specifier (*break ice*)
- In 3 examples there is a different specifier (*a lot of* (2), *some of the*)
- 3 examples have a modifier (*diplomatic* (2), *new*)
- 3 examples involve the compound noun *post-Cold War*

10 examples are passive:

- the ice is broken (2)
- the ice was broken (2)
- the ice has been broken (3)
- But the ice was later broken at the following plenary session [...]
- the ice was quickly broken with the hunters.
- Catholic politicians were glad the ice was being broken [...]

In 4 inchoative examples *the ice* is the subject: *the ice has broken*.
12 examples are unusual in other ways:

- [...] the breaking of ice between Palestinians and Israelis [...]
- We have made the ice break [...]
- Christopher, evidently intent on getting ice properly broken, walked around the table and whispered something to the three.
- [...] however we got the ice broke
- There was certainly ice to be broken
- The ice began to break when Beijing sent human rights activist Harry Wu home
- The ice began breaking this spring, when British officials prodded BT [...]
- Arabs hoped the ice that had shrouded Middle East peace talks since the Israeli election had been broken.
- Chirac and Balladur break bread, but not the ice
- How dare you break up the ice in your town?
- [...] we're going to see the ice breaking up even more [...]
- [...] were a symbolic break in the ice.

level the playing field

79% of the occurrences of this idiom, which roughly means 'equalize the situation', are in their canonical form: 349 out of 443 show no variation other than inflection of *level*.[38]

There are 94 non-canonical examples of this idiom, which represents 21% of the occurrences in the corpus.

- In 2 examples *playing fields* is plural
- In 8 examples there is a different specifier (*a* (6), *this* (2))
- 35 examples have a modifier (*military* (7), *political* (6), *electoral* (4), *financial* (3), *competitive* (3), *professional* (2), *digital* (2), *academic* (2), *educational*, *legal*, *economic*, *industrial*, *global*, *democratic*)
- 3 examples have a different specifier and a modifier (*Mexico's political*, *the country's electoral*, *a new electoral*)
- 15 examples involve a compound noun (*the export playing field* (2), *the wage playing field* (2), *the trade playing field* (2), *the mortgage playing field*, *the employment playing field*, *the business playing field*, *the goodwill playing field*, *the benefits playing field*, *the parts playing field*, *the international-trade playing field*, *the criminal-justice playing field*, *the land-use playing field*)

15 examples are passive, and some of them have various other interesting properties:

- The playing field has been leveled through redistricting [...]
- Now that the playing field has been leveled [...]
- [...] the professional playing field has not merely been leveled [...]
- [...] not only has life's playing field been leveled [...]
- Has the playing field been leveled?
- The playing field needs to be leveled [...] (3)
- [...] the playing field must be leveled [...] (2)
- The playing field cannot be leveled [...]
- [...] the playing field should be leveled [...]

[38]Note that there is a related expression with the adjective *level*, as in the idiomatic NP *a level playing field*. There are 695 occurrences of it in the corpus, 474 (68%) of which are canonical.

- [...] the playing field ought to be leveled [...]
- We have waited a long time for the competitive playing field to be leveled [...]
- [...] we want the playing field leveled so we have an even chance [...]

16 examples are unusual in other ways:

- leveling of the playing field (7)
- a leveling of the political playing field
- a kind of leveling of the squirrel playing field
- an almost instantaneous leveling of the league's playing field.
- The playing field leveled
- [...] unless the playing field starts levelling [...]
- [...] the playing field levels out [...]
- Gary Nelson has leveled up the playing field in NASCAR [...]
- We have a responsibility to support efforts to raise the playing field even as we work to level it [...]
- Asia's leveled playing field has produced a far larger middle class.

lead the field

57% of the occurrences of this idiom, which roughly means 'be ahead of the competition', are in their canonical form: 96 out of 169 show no variation other than inflection of *lead*.

There are 73 non-canonical examples of this idiom, which represents 43% of the occurrences in the corpus.

- In 5 examples there is no specifier (*lead field*) and in 2 of these there is the modifier *Republican*
- In 24 examples there is a different specifier: 20 examples are of the form *lead a field* (followed by *of* or a relative clause), and 4 examples have other specifiers (*the rest of the* (3), *this year's*)
- 21 examples have a modifier (*Republican* (15), *Republican presidential* (4), *crowded*, *GOP presidential*)

- 8 examples have a different specifier and a modifier (*a crowded* (5), *a big*, *a limited*, *a Republican*)
- 15 examples involve a compound noun (*the GOP field* (6), *the Republican presidential nomination field*, *the skimpy musical field*, *the eight-man field*, *the existing GOP field*, *the Duma field*, *the seven-candidate field*, *a crowded GOP field*, *a five-candidate field*, *a three-man field*)

Note that *field* with this meaning can also be used in other expressions like *trail the Republican presidential field*, *lead among the field* and *lead in the field*. In fact it is part of the more general *race* metaphor for elections (*run for president*, *Congressional race*). However, I am not sure whether *field* can really have the meaning 'competition' or 'competitors' in general (?*the field for this job search is big/fierce*).[39] If *field* can be used this way, that suggests that this expression does not fit my definition of idiom. This seems to be the case for at least some speakers.

take a back seat (to)

94% of the occurrences of this idiom, which roughly means 'take a subsidiary position (to)', are in their canonical form: 658 out of 700 show no variation other than inflection of *take*. Note that I counted the alternative spelling of *backseat* (used 57 times) in the same way as *back seat*. This assumes that these words are stressed in the same place and are just spelling variants.[40] If occurrences of *backseat* were counted as varied, the percentage of canonical occurrences would be 86%.
There are 42 non-canonical examples of this idiom, which represents 6% of the occurrences in the corpus.

- In 1 example *seats* is plural
- In 6 examples there is no specifier (*take back seat*)

[39] According to Merriam-Webster's Collegiate Dictionary *field* can refer to competitors in a sports activity only, and I did not count sports uses as instances of this idiom. The Collins Cobuild Dictionary says that 'field' is used in expressions such as *hold the field* and *lead the field* when there is competition in general, but does not say whether or not the list of such expressions is limited to a few conventionalized ones.

[40] I asked several speakers, and some stressed *back* while others stressed *seat*, but none stressed the two spelling variants differently from each other.

- In 18 examples there is a different specifier (*the* (11), *very much a* (2), *no*, *a bit of a*, *more of a*, *somewhat of a*, *something of a*)
- 13 examples have a modifier (*distant* (3), *temporary* (2), *undeserving*, *political*, *understandable*, *cultural*, *corporate*, *sheepish*, *unwarranted*, *socially accepted*)
- 1 example involves a compound noun (*a public-relations back seat*)

3 examples are unusual in other ways, including two occurrences across relative clause boundaries.

- [...] the back seat President Boris N. Yeltsin has taken to Chernomyrdin in the crisis suggested he, too, might have preferred a resort to brute force to resolve the stand-off.
- He mourns the backseat that books have taken to movies [...]
- It is a fine line we walk as we try to teach our daughters to take the reins of the world, not a back seat.

Note that 60% of the corpus occurrences are of the form *take a back seat to*. One may wonder whether this idiom may be nondecomposable, as in many examples it seems to mean simply 'defer to':

- Usually outspoken House Speaker Newt Gingrich says he will take a back seat to Bob Dole [...]
- [...] the leader of the alliance had to take a back seat to its junior partners

Although these examples are also consistent with the decomposable interpretation, there may be some speakers who think of the meaning this way. This would explain why this idiom exhibits an unusually low degree of variation. Furthermore, most of the modifiers are best thought of as external modification, e.g. *take an unwarranted back seat* means *unwarrantedly take a subsidiary position*, not ?*take an unwarranted subsidiary position*, and similarly for *undeserving* and *understandable*. Some other modifiers make more sense as immediate modifiers of 'position' than of 'subsidiary position' (*subsidiary cultural position* vs. ?*cultural subsidiary position*), which is not possible as nothing in the idiom corresponds directly to 'position'. In any case, the external modification analysis works for them as well: *temporarily take a subsidiary position*, *take a subsidiary position what culture is concerned*, etc.

Note also that the Collins COBUILD idioms dictionary claims that this idiom has two meanings, depending on whether the subject is a volitional agent:

> **take a back seat:1**
>
> If you **take a back seat**, you allow other people to have all the power, importance, or responsibility.
>
> **take a back seat:2**
>
> If one thing **takes a back seat** to another, people give the first thing less attention because they think that it is less important or less interesting than the other thing.

The second meaning is by no means rare in the corpus:

(124) a. Even within the EPA, the drinking water program takes a back seat.
b. Even the government's fight against inflation took a back seat.
c. Other priorities took a back seat.

While both of these meanings are compatible with the paraphrase *take a subsidiary position*, it is possible that some speakers think of this as two meanings of *take*, or have somewhat different interpretations like *defer to* vs. *be considered less important*. Both of these are non-decomposable in that their parts do not correspond to the parts of the idiom, which would explain the higher percentage of canonical forms.

3.5.2 Non-Decomposable Idioms

hit home

100% of the occurrences of this idiom, which roughly means 'register painfully', 'become painfully apparent', are in their canonical form: 276 out of 276 show no variation other than inflection of *hit*.

Note that I consider *hit close to home* to be a separate expression.[41] It does not seem possible to have one representation that captures what the two idioms have in

[41] If it were not treated as a separate expression or as a conventionalized variant, but as productive variation, then 87% of the occurrences of *hit home* would be canonical. Variations of *hit close to home* include *too close*, *so close*, *very close*, *closer*, and *closest*. Note that this expression's structure allows for more potential variation, and it would not have been in my list of V+NP idioms. In fact,

common, as the syntactic relationship between *home* and *hit* is not the same in both cases. I do not think the two idioms mean exactly the same thing either, although the difference is not very large in some examples. But *hit close to home* roughly means 'affect one personally', as can be seen from the examples below:

- Who better to talk about getting off drugs than someone who's seen it hit close to home?
- When a crisis hits close to home and disrupts regular and comforting routines, children become more anxious [...]
- Rahe Mulligan, 21, says she is more concerned about the environment, AIDS and other issues that "hit close to home."
- [...] issues such as welfare reform hit close to home for some students.

In contrast, *hit home* means 'register painfully' or 'become painfully apparent', and it is often used with nouns like *reality*, *facts*, and *truth*:

- The reality of my addiction didn't hit home until last week.
- But the cold facts hit home Wednesday.
- Then Sunday, the truth of it began to hit home.
- When the ugly truth hits home about the expense and inconvenience of accommodations, a lot may decide to eat their ticket losses and watch the Games on NBC.

For example *the reality of my addiction hit home* means 'the reality of my addiction registered painfully' or 'the reality of my addiction became painfully apparent', not *the reality of my addiction affected me personally* (which it presumably already did before the speaker became aware of it).

The two idioms are also not used the same way syntactically. For example, there is no occurrence of *hit close to home with*, while *hit home with* 'register painfully with' is quite frequently used (in 36 examples):

- Because few Mexicans can afford a car, the rise in gasoline prices hasn't hit home with most people.

it is probably not an idiom but a collocation. It is possible that *hit home* should not have been on the list, either, as *home* seems to be directional and should perhaps be analyzed as a PP rather than an NP.

- He said the allegations would hit home with PC owners.
- The message appeared to hit home with many in the crowd [...]

speak volumes

100%[42] of the occurrences of this idiom, which roughly means 'reveal a lot', are in their canonical form: 428 out of 430 show no variation other than inflection of *speak*. There are 2 non-canonical examples of this idiom, one of which is presumably a typo, and the other of which I would probably want to call 'word play'.

- How the university celebrates itself on its 250 year speaks volume about what Princeton is.
- The rescued Isabella sings not a note, but her total nudity speaks, if not volumes, at least a couple of pungent paragraphs.

close ranks

94% of the occurrences of this idiom, which roughly means 'be supportive', 'be united', are in their canonical form: 363 out of 386 show no variation other than inflection of *close*.

There are 23 non-canonical examples of this idiom, which represents 6% of the occurrences in the corpus.

- In 1 example *rank* is singular
- In 7 examples there is a different specifier (*the* (3), *their* (2), *its* (2))
- 4 examples have a modifier (*Arab* (2), *Democratic*, *Palestinian*)
- 1 example has a different specifier and a modifier (*his party's divided ranks*)
- 7 examples are of the form *closing of ranks*
- 3 examples are of the form *closing of the ranks*

There are some possible V+NP paraphrases for this idiom, e.g. 'increase unity' or 'reduce divisions'. But the idiom is not semantically decomposable because there is no mapping from these paraphrases to parts of the idiom, i.e. *ranks* does not correspond to *unity* or *divisions*, but to the group that is being united.

[42]The 100% figure is the result of using rounding the same way as for the other idioms. The actual percentage is 99.53%.

look the other way

98% of the occurrences of this idiom, which roughly means 'deliberately ignore (it)', 'tolerate (it)', are in their canonical form: 453 out of 462 show no variation other than inflection of *look*.

There are 9 non-canonical examples of this idiom, which represents 2% of the occurrences in the corpus. They are of the form *look the other way at*.

- [...] it is at times difficult to accept Greg's ability to look the other way at his own transgressions [...]
- [...] the United States at best looked the other way at corruption in the ruling anti-Communist parties [...]
- [...] looking the other way at such illegal conversions has eased a catastrophic housing crisis [...]

These examples must be considered varied although the string is the same, because the prepositional phrases headed by *at* are complements of *look*. It is interesting that the choice of preposition seems to be inspired by literal *look*,[43] even though these examples clearly cannot be interpreted literally (*they looked the other way at his transgressions* does not mean 'they looked differently at his transgressions' but 'they closed their eyes with respect to his transgressions', i.e. 'they deliberately ignored his transgressions'). That is, it looks like these prepositional phrases are used in a similar way as in *they were surprised at his transgressions*. This is a way of integrating a complement for 'ignore' into the structure of this idiom, which does not usually allow for a complement.

Note that there are also 7 occurrences of *look the other way on*. I counted these as canonical because the prepositional phrases can be fronted.

- I could look the other way on these issues [...]
- NATO looks the other way on Bosnian war criminals [...]
- While the administration has looked the other way on drugs [...]

[43] An interesting question for further research is whether this is systematically the case for other verbs as well.

Another interesting thing to note about this idiom is that it is semantically non-decomposable although it is not an idiom of decoding (Makkai 1972:57). That is, a speaker who has never heard this idiom before can probably figure out what it means and see the figurative connection between the idiom and its literal meaning. So transparency is not a sufficient condition for semantic decomposability in the sense in which this concept is used in the dissertation.

clear the air

93% of the occurrences of this idiom, which roughly means 'address the misgivings/misunderstandings/hard feelings',[44] are in their canonical form: 273 out of 292 show no variation other than inflection of *clear*.
There are 19 non-canonical examples of this idiom, which represents 7% of the occurrences in the corpus.

- In 4 examples there is no specifier (*clear air*)
- 3 examples have a modifier (*any poisonous political air*, *the general air*, *the political air*)
- 1 example involves a compound noun (*the campaign air*)
- 4 examples involve *clearing of (the) air*, including 1 *clearing of the political air*
- 2 examples involve *clear up* (*clear up the air* and *the air cleared up*)

5 examples are unusual in other ways, including passives:

- I don't have a shred of doubt that the air can be cleared
- A lot of air has been cleared
- Now the air has been cleared
- It was good the air was cleared.
- get the air cleared

It is clear that *the air* does not correspond to 'misgivings', 'misunderstandings', or 'hard feelings' in this idiom. Instead, *the air* corresponds to the whole situation (*clear the air of misgivings* = 'eliminate misgivings from the situation'), so this idiom is non-decomposable. Modifiers like *clear the political air* and *clear the campaign*

[44] I did not count occurrences of literal air clearing, e.g. after a fire, as instances of this idiom.

air can probably be thought of as external modifiers ('address the misgivings what politics are concerned', 'address the misgivings what the campaign is concerned' etc.). However, the passive examples suggest that the people who used them may have had a decomposable interpretation of the idiom.

bite the bullet

95% of the occurrences of this idiom, which roughly means 'accept the (difficult) situation' or 'face reality', are in their canonical form: 188 out of 198 show no variation other than inflection of *bite*.
There are 10 non-canonical examples of this idiom, which represents 5% of the occurrences in the corpus.

- In 1 example there is no specifier and the noun is plural (*bite bullets*)
- In 1 example there is no specifier (*bite bullet*)
- In 3 examples there is a different specifier (*that*, *this*, *a*)
- 2 examples have a modifier (*financial*, *fiscal*)
- 1 example involves a compound noun (*the reform bullet*)
- 1 example is *bite the bullet of reality*

1 example is passive (and also involves a compound noun):

- But it increases the odds that if the COLA bullet is bitten, retired feds won't be the only group with tooth-marks on their hides.

Note that 89 (=45%) of the examples are followed by *and*. This construction serves the function of further specifying what constitutes *accepting the situation*.

- A lot of investors don't want to bite the bullet and take the loss [...]
- [...] you have to bite the bullet and talk with a tech-support rep.
- [...] bite the bullet and study the manual [...]
- Greenspan has called upon Congress to bite the bullet and cut the deficit [...]

This idiom was historically a metaphor for 'diverting attention from the pain of an operation'. It clearly does not have that meaning any more, but the figurative meaning still involves having to take an unpleasant or undesired action. In this metaphor

the bullet does not correspond to the whole situation, so the idiom is not decomposable. Furthermore, it is hard to think of the verb *bite* as metaphorical for 'accept'. With verbs like idiomatic *spill* the metaphoric connection to 'reveal' can be seen even in the absence of *beans*. The same is true for other verbs like *turn* and 'reverse', *break* and 'escape', *level* and 'equalize', etc. It is possible to use most of these verbs with these metaphoric meanings in other idiomatic and even non-idiomatic utterances (*spill the secrets*), while the same is not true for *bite* (**bite the compromise*). This idiom has also been discussed as an example of a non-decomposable idiom in the literature (e.g., by Nunberg et al. (1994)).

Note that the examples *bite the financial bullet* and *bite the reform bullet* can be given interpretations of external modification ('accept the situation financially', 'accept the situation what the reform is concerned'). However, the person who used the passive example probably thought of this idiom as decomposable.

3.5.3 Summary

A summary of the data from the study of randomly selected V+NP idioms is given in Table 3.2. If one leaves out the two items that are probably collocations as opposed to idioms for many speakers (*pay dividends* and *lead the field*), the percentage for decomposable idioms is 75% instead of 73%. This is quite similar to the 80% figure I found in Section 3.2 for the decomposable idioms from the literature. Similarly, the figure of 97% for non-decomposable idioms from the random sample is similar to the 96% figure for the non-decomposable idioms from the literature.

3.6 Collocations

bear the brunt of

There are 749 occurrences of this collocation in the corpus. 704 (94%) of them are canonical in the sense of being the string *bear the brunt* modulo inflection of *bear*, and 675 (90%) are canonical if the stricter criterion of being *bear the brunt of* is used.

This expression is clearly not an idiom in that its meaning is predictable and

	Total # of Tokens	**% Canonical**	**% Variation**
Decomposable Idioms			
turn the tables	518	69%	31%
call the shots	589	79%	21%
deliver the goods	176	84%	16%
lose face	137	85%	15%
make waves	243	77%	23%
run the show	368	78%	22%
pay dividends	418	52%	48%
sound the death knell	110	66%	34%
break the mold	168	61%	39%
lose ground	2350	70%	30%
strike a chord	688	57%	43%
rear its head	128	74%	26%
break the ice	183	79%	21%
level the playing field	443	79%	21%
lead the field	169	57%	43%
take a back seat	700	94%	6%
Average:		**73%**	27%
Non-Decomposable Idioms			
hit home	276	100%	0%
speak volumes	430	100%	0%
close ranks	386	94%	6%
look the other way	462	98%	2%
clear the air	292	93%	7%
bite the bullet	198	95%	5%
Average:		**97%**	3%

Table 3.2: Results from the Study of Random V+NP Idioms

compositional, and all its parts occur with the same meanings in other expressions (e.g. *take the brunt of* (156 occurrences in the corpus), *feel the brunt of* (68), *suffer the brunt of* (26), *get the brunt of* (19), *absorb the brunt of* (10), *face the brunt of* (9), *receive the brunt of* (8), *shoulder the brunt of* (5), etc.), for a total of 309 occurrences with similar verbs. There are also 172 occurrences with no such verb, for example:

(125) a. The brunt of the warm winter storm swept south.
b. Asia may be spared the brunt of the cuts.
c. But Western nations are expected to provide the brunt of troops and resources for any multinational force.

10% of the occurrences of this collocation are non-canonical if one counts a missing *of*-complement as non-canonical, 6% otherwise.

- 5 examples are headlines with a missing specifier (e.g. *Biotechs bear brunt of jittery market*)
- 9 examples exhibit a variation in the specifier (*much of the brunt of*, *more of the brunt of*, *most of the brunt of*, *most of the brunt*, *such a brunt*, *its brunt*, *that brunt*)
- 16 examples are modified (*the great brunt of*, *the financial brunt of*, *the double brunt of*, *the full brunt of*, *the major brunt of*, *the entire brunt of*, *the full year-over-year brunt of*, *a disproportionate brunt of*, *the confusing brunt of*, *the biggest brunt of*, *the biggest brunt*, *the ultimate brunt*, *the psychological if not the physical brunt*)

15 examples are passive:

(126) a. He told Reuters the brunt of the cyclone's fury may be borne by Tamil Nadu rather than Andhra Pradesh.
b. Most Israelis are well aware that the brunt of the pain is being borne by Lebanese civilians and the government in Beirut.
c. Since the midsummer correction, the brunt of which was borne by small-cap technology stocks, investors have taken a different approach [...]
d. The brunt of losses will be borne by those member credit unions.

e. The brunt of the cost of the new system may be borne by the meat and poultry processing industry.

1 example involves raising in addition to passive:

(127) The brunt of the cuts is likely to be borne by 12,000 workers [...]

give (someone) the green light

I studied this expression because I thought it was an idiom with an 'open slot'. However, I searched for *green light* because I thought it was infrequent enough to do so, and it turns out that only 62% of the 1479 occurrences of non-literal *green light* are with *give*. While many of the other occurrences are of the form *get the green light*, which is potentially another fixed expression, in 27% of the examples other verbs are used, such as *receive*, *await*, *wait for*, *expect*, *seek*, and *need*. Some of these verbs fit the metaphor of giving and receiving but are clearly not conventionalized. Other verbs used with *green light*, such as *seen as*, do not fit the metaphor of giving and receiving, and in fact in many examples there is no such verb present at all. So *green light* can clearly have its idiomatic meaning when it occurs by itself.

(128) a. A green light from the panel would make approval all but certain.
b. This is not a green light for everyone to show up in the emergency room
c. Analysts called the approval a major green light for both companies [...]
d. [...] the action amounts to a green light.
e. The proposal "may be seen as a green light to demolish what little children's educational programming still appears on commercial television [...]

However, *give (someone) the green light* is nevertheless conventionalized. As it does not fit my definition of idiom it has to be considered a collocation instead. There are 922 occurrences of *give ... green light*. 366 are of the form *give the green light* and 198 are of the form *give NP the green light*. Together these two forms account for 61% of the occurrences of this collocation. Note that because of the possibility of this alternation, there is no one canonical form of this expression, the most frequent single form accounting for only 40% of the data. It would be interesting to see whether a similar pattern holds for real idioms with open slots, or whether it is due to the fact

that *green light* can exist independently, and can therefore combine freely with both lexical entries for *give*.

- In 27 examples there is no specifier (*give green light* (19), *give NP green light* (7))
- 170 examples are of the form *give a green light*
- 89 examples are of the form *give NP a green light*
- 10 examples involve other specifiers (*one's* (7), *a kind of*, *a lot of*, *all but a*)
- 13 involve *the* and a modifier (*official* (4), *final* (3), *formal* (2), *initial*, *secret*, *needed*, *alleged*)
- 33 examples involve *a* and a modifier (*virtual* (6), *final* (2), *indirect* (2), *formal*, *rhetorical*, *international*, *legal*, *Syrian*, *vital*, *tacit*, *secretive*, *medical*, *tentative*, *temporary*, *secret*, *preliminary*, *diplomatic*, *total*, *positive*, *rapid*, *medium*, *so-called*, *long-sought*, *very clear*, *very public*, *indirect or direct*, *series*)
- 1 example involves another specifier and a modifier (*its required*)

8 examples are passive:

- No final green light has been given for the Finnish team to proceed with their work.
- [...] to see if any "green light" was given for an attack.
- [...] even if a green light is given by his union.
- [...] after the green light has been given to us [...]
- Two weeks after the green light was given [...]
- [...] before the green light was given.
- Once it is signed, the green light will be given [...]
- [...] when the green light is given in Washington [...]

2 examples involve subordinate clauses:

- The senator herself has left a long trail of hints: [...] the green light she gave to other candidates eager to start campaigns to succeed her.
- Doug Fabian says his decision to only half-obey the green light given by the Fabian index isn't a direct response to the system's recent laggard performance

5 examples are varied in other ways:

- [...] after initially giving the greenest of green lights to Israel [...]
- [...] the United States had given the Croatians 'a green or amber light' to proceed [...]
- [...] giving Israel the green or yellow light [...]
- [...] basic principles that give Jason Kidd the ball and his teammates the green light to shoot at any time [...]
- [...] former coach Chris Gobrecht gave Redd an average of 28 minutes and the green light.

play fast and loose

I thought this was an interesting idiom to study because *fast and loose* is not a complement of *play*. However, it turns out that *play* is not an essential part of the idiom, as *fast and loose* occurs 19 times without *play*. So the actual idiom is *fast and loose*, and *play fast and loose* has to be considered a collocation.
89% of the occurrences of this collocation are in their canonical form: 95 out of 107 show no variation other than inflection of *play*.
12 of the occurrences of this collocation are non-canonical:

- 5 examples involve event modification: (*play a bit fast and loose*, *play a little fast and loose*, *play so fast and loose*, *play too fast and loose*)
- 4 examples involve inserted complements (*play it fast and loose*, *play politics fast and loose*, *play the writing game fast and loose*, *play things fast and loose*)
- Other variation includes *play as fast and loose ... as* and *play fast or loose*

Note that *play the writing game fast and loose* suggests that *play* must have the same meaning in *play the game*, perhaps a literal meaning.

caught between a rock and a hard place

I thought this was an interesting idiom to study as it involves more than head-argument relationships. However, it turns out that *caught* is an optional element, so that the real idiom is *between a rock and a hard place*, and *caught between a rock and a hard place* has to be considered a collocation.

4 out of 7 examples of this collocation are canonical. The non-canonical occurrences of this collocation are:

(129) a. He is between a rock and a hard place
b. This put Ceroils "between a rock and a hard place," he said [...]
c. Palestinian President Arafat, caught between the rock of Israeli policy and the hard place of a population in deepening need, had pressed Israel to relax the closure ahead of the summit.

3.7 Constructions

(*be*) *nothing if not*

I studied the 190 occurrences of this construction in the New York Times corpus that is part of the North American News Text corpus.
An inflected form of *be* followed immediately by *nothing if not* is involved 93% of the time. Some typical examples are:

(130) a. Hugh Grant is nothing if not charming.
b. But baseball fans are nothing if not forgiving.

However, there are two examples where *also* intervenes between *be* and *nothing*, so it's not a completely fixed phrase:

(131) a. Dole is also nothing if not protective of his own pride.
b. He is also nothing if not self-confident.

And there are 12 examples that do not involve the verb *be*:

(132) a. For the rare visitor, he has nothing if not time.
b. These characters have nothing if not lots of spunk [...]
c. Last week, Barry's setback did nothing if not magnify the view of Washington as a city on the brink of a breakdown.
d. By the numbers, the town represents nothing if not an aberration.
e. Mag knows nothing if not herself.
f. The Bushes inherit nothing if not the competitive spirit.

g. [. . .] the first images of this video [. . .] convey nothing if not sublime delicacy.
h. America was founded on nothing if not the principle of individual freedom and responsibility.
i. And Rugby stands for nothing if not tradition.
j. For a book of Venice - famous for nothing if not its canals [. . .]
k. [. . .] but that's the new GOP for you: nothing if not alert against the unseemly petitions of the poorly off.

One might think that the literal meaning plus some Gricean rule of interpretation would give the right meaning: (X is nothing if X is not Y) & (X is not nothing) → (X is Y). However, *Hugh Grant is nothing if not charming* does not simply mean *Hugh Grant is charming*. Instead it has an intensified meaning, asserting a more emphatic claim like 'Hugh Grant is certainly charming' or 'Hugh Grant is really charming'.

Also note that even if the correct interpretation could be derived, the form in which it is expressed is still conventionalized. Otherwise one would also expect examples of the form **Hugh Grant is nobody if not charming* and #*Hugh Grant is nothing if he is not charming*. The latter would be expected in analogy to *You are likely to fall if you are not careful*, where the shorter form ??*You are likely to fall if not careful* is not as acceptable.

It is not possible to associate this meaning with any of the words making up this construction. While for examples involving a form of *be*, *certainly* or *really* can just replace *nothing if not* in the paraphrase, that is not true for the other examples:

(133) a. He has nothing if not time.
b.*He has certainly time.
c.*He has really time.

Yet the construction has the same meaning in (133a) and other examples without *be*: *He certainly has time* or *He really has time*. It does not seem possible to get *really* to have scope over the verb in a non-constructional analysis.

3.8 Comparison with Non-Idioms

In contrast to the high level of canonicity observed in idioms, there is no such phenomenon with semantically similar non-idiomatic expressions. This can be seen as a baseline that shows there is no semantic or pragmatic reason why, e.g., an unmodified definite plural is so frequent in *spill the beans* (87%). The form *reveal the secrets* accounts only for 1% of the sentences involving 'reveal' and 'secret',[45] and there is no other canonical form, either—the most frequent form is *reveal secrets* (7%). Modifications and syntactic variations are much more frequent than with idioms—they account for 87% of the corpus occurrences, i.e. only 13% of the occurrences are of the form *reveal (a/the) secret(s)*. And even the definite article for *secret(s)* is not particularly frequent—it is part of only 18% of the corpus examples, including the modified and syntactically varied ones. One might think that this shows that *the* does not have its literal meaning in *spill the beans*, but this does not seem to be the case. Fellbaum (1993:285) shows that *the beans* like *the secrets* refers to secrets whose existence is known to the discourse participants.

For many idioms it is hard to find a paraphrase that closely matches them both syntactically and semantically. However, it is still striking that the difference between idiom and non-idiom is so large even when a reasonable paraphrase exists. Furthermore, it is interesting in and of itself that that non-idioms in general have nothing resembling a canonical form. If the high level of canonicity in idioms were not due to conventionalization but instead due to general semantic or pragmatic principles, one might expect these same principles to be at work with non-idioms. Note also that for non-idiomatic expressions, there is not just more variation in terms of the percentage of varied examples, but there are also more types of variation observed.

reveal secrets

Out of 387 sentences about the 'revealing' of 'secrets', the form that matches *spill the beans* in number and definiteness, i.e. *reveal the secrets*, occurs only 5 times,

[45] Like with the idioms, I counted all examples involving the verb *reveal* and the noun *secret* in the right semantic relationship as instances of this expression.

accounting for 1% of the data. The most frequent form is *reveal secrets*. But there are only 27 occurrences of it, accounting for only 7% of the data, and therefore it cannot be called a canonical form. Even if one includes the 10 additional examples of *reveal secrets of NP* (like *reveal secrets of the Inca culture*), the total accounts for only 10% of the data. There is no other canonical form either: *reveal the secret* (8) and *reveal a secret* (9) are even less frequent.

93% of the 387 instances of this expression are varied. In 87% of them the complement is not of the form 'secrets', 'the secrets', 'a secret', or 'the secrets', and even if one counts '(the) secret(s) of' as 'plain' there is still 79% variation.

- 48 examples involve other specifiers, such as possessive pronouns, and *any*, *no*, *many*, *few*, *some*, *enough*, *a number of*, *all of his*, *the drivers'*, *many of Santa's*
- 25 examples involve modifiers (*embarrassing*, *military*, *Russian*, *personal*, *dark*, *guilty*, *deep*, *unguessed*, *industrial and political*, *perhaps unimaginable*, *potentially damaging*, *valuable corporate*)
- 47 examples involve both specifiers and modifiers (*his coaching secrets*, *her long-held secret*, *a dark secret*, *Mitterrand's long-held secret*, *Ozzie's little secrets*, *some deep, dark secret*, *a variety of troublesome secrets*, *his dark secret*, *the Soviet Union's darkest secrets*, *the Jewish state's nuclear secrets*, *any real secrets*, *some of the CIA's deepest secrets*, *their most intimate secrets*, *a closely guarded secret*, *doctors' dirty little secrets*, *one of the Cold War's darkest secrets*, *Italy's pretty little secret*, *many of America's most vital cold war secrets*, *some of the royal family's deepest secrets*)
- 56 examples involve plain compound nouns (*state secrets* (32), *trade secrets*, *government secrets*, *agency secrets*, *nuclear arms secrets*, *family secrets*, *company secrets*, *love secrets*, *personality secrets*, *bedroom secrets*)
- 11 examples involve compound nouns plus specifiers (*a family secret*, *Lassie's beauty secrets*, *all our family secrets*, *some family secrets*, *any family secrets*, *their trade secrets*, *a state secret*)
- 4 examples involve compound nouns plus modifiers (*"compromising" Kremlin secrets*, *embarrassing agency secrets*, *explicit bedroom secrets*, *valuable trade secrets*)

- 6 examples involve compound nouns plus specifiers and modifiers (*some unbecoming pageant secrets*, *a long-held family secret*, *a dirty family secret*, *one of the country's best kept business secrets*, *a dark family secret*, *their software trade secrets*)
- 14 examples involve variations of 'secrets of': 6 involving specifiers (*one of the secrets of*, *some of the secrets of*, *some secrets of*), 3 involving modifiers (*other secrets of*, *alleged intimate secrets of*), and 5 involving both specifiers and modifiers (*the deepest and sometimes most shocking secrets of*, *the dirty secrets of*, *the horrid secrets of*, *the inner secrets of*, *the innermost secrets of*)

36 of the examples are passives. In 30 of those there is some sort of modification involved as well.

(134) a. And so our dirty little secret was revealed.
b. Only recently have the secrets of Movile been revealed to the world.

In addition there are 8 adjectival participles:

(135) The story, which involves a slowly revealed secret on the young woman's part, will probably not hold the interest of even the most romantic pre-teen-age girls.

In 37 examples 'reveal' and 'secret' occur across multiple clauses. 21 out of these are passivized as well, and some involve raising.

(136) a. In short, Britain could talk of nothing else but the secrets Diana was likely to reveal in the program taped at apartments in Kensington Palace.
b. Some secrets really were meant to be revealed.
c. "They were carefully guarded secrets, revealed only to the initiates, not outsiders," Ulansey said.
d. The secret, which you can see her revealing Thursday night to friends, schoolmates and the whole town, is that she was born with AIDS.

There are 13 other examples:

(137) a. Percy has secrets to reveal, and she's not the only one.
b. Who, exactly, are these "salty old women" whose secrets Bird professes to reveal?

c. [. . .] decide when a secret is no longer too delicate to reveal.
d. [. . .] who knows a mob secret but fears reprisals if he reveals it to government prosecutors.
e. "A policy on national secrets must be identified so that whatever is not 'classified' can be revealed to those who can justify their need."
f. What secrets do those Deepak Chopra self-help tomes reveal about you?
g. "What secrets does he know now that he hasn't revealed to the newspapers?" said Feldman.
h. "What he revealed was not state secrets, but [. . .]
i. More than one person landed in jail for revealing something about them, one of the cold war's top secrets.
j. Shawn, however, reveals the real truth, which is a "secret" that requires adult assistance.
k. The third night reveals, for the first time, some - but not all - of Nilsson's photographic equipment and secrets.

Many of these types of variation are not observed with idioms.

divulge information

Out of 138 sentences involving the 'divulging' of 'information', 42 are of the form 'divulge information'. This corresponds to 31%. This somewhat higher percentage is partly due to the fact that varying number is not an option. The expression may also be somewhat conventionalized, as *divulge* does not collocate with many nouns. Even this percentage does not come close to the level of canonicity observed in idioms. The remaining 69% of the occurrences of this expression are varied.

- 3 examples involve *the information*
- 23 examples involve other specifiers (*any*, *any more*, *such*, *the same*, *that*, *more*, *as much*, *any such*, *all*, *only selected pieces of*, *one or two items of*)
- 32 examples involve modifiers (*detailed*, *damaging*, *sketchy*, *little*, *military*, *restricted*, *confidential*, *proprietary*, *critical*, *personal*, *inside*, *classified*, *financial*, *sensitive*, *internal*, *certain*, *corporate*, *classified military*, *sensitive and important*, *embarrassing or secret*)

- 7 examples involve specifiers and modifiers (*any top-secret*, *any other financial*, *the new*, *more personal*, *a client's confidential*, *such highly confidential*, *as little personal information as possible*)
- 8 examples involve compound nouns (*insider information*, *key information*, *intelligence information*, *confidential business information*, *other critical law-enforcement information*, *confidential customer information*, *employment and financial information*, *the patent information*)

7 examples are passivized:

(138) a. [...] confidential information [...] is not divulged to AT&T's own cellular unit.

b. Ickes gave testimony [...] about information allegedly divulged to the White House by the Resolution Trust Corp. [...]

c. "I have expressed real concern about information being divulged by whomever is divulging it."

d. [...] they fear that personal information will be divulged without discretion.

e. [...] senior officials also discussed how Washington can use secret information being divulged by the defectors about Iraq's clandestine production of weapons [...]

f. O.J. Simpson lead defense attorney Johnnie Cochran Jr., debating the merits of information divulged on a sports interview TV show [...]

g. Information divulged from such global networks should be trustworthy [...]

In 7 examples the parts of the expression occur across clause boundaries, such as relative clauses:

(139) a. [...] Nuccio acknowledges that the information he divulged was supposed to be kept secret.

b. The company also is not providing any social and economic information customers divulge to the telephone company.

c. [...] White's posting was based upon confidential information that the girl divulged in a private group discussion [...]

d. This is competitive information they do not want to divulge [...]

In 3 examples the expression occurs across clause boundaries and involves pronouns:

(140) a. [...] when he pressed the school and the coaching staff for information on his accuser, they refused to divulge it.

b. He must decide if Savage's information is so crucial to Simpson's defense that she must divulge it despite the shield.

c. The Clinton administration is proposing [...] that the FBI be allowed to provide information in such cases to a judge without divulging it to the person it wants expelled.

And 5 examples are varied in other ways:

(141) a. [...] he has more relevant information than he divulged in court.

b. Presidential staffer Harold M. Ickes [...] contradicted Altman on what information the Treasury deputy had divulged to the White House.

c. Kyaw Ba said whether Khun Sa was charged and prosecuted depended on what information he was willing to divulge [...]

d. Others agonize over how much patient information they must divulge.

e. [...] they would divulge specific trade secrets and confidential information

divulge secrets

Out of 73 sentences involving the 'divulging' of 'secrets', 9 are of the form 'divulge secrets'. This corresponds to 12%. The remaining 88% of the occurrences of this expression are varied.

- In 8 examples *secret* is singular (*the secret* (3), *a secret*, *his secret* (2), *a sordid secret*, *the great secret*)
- 4 examples involve *the secrets*
- 8 examples involve other specifiers (*no* (2), *many*, *its* (2), *all the*, *others'*, *only such*)
- 4 examples involve modifiers (*military* (2), *competitive*, *personal*)
- 2 examples involve specifiers and modifiers (*the intimate*, *the dark*)

- 28 examples involve compound nouns (*state secrets* (10), *company secrets* (5), *trade secrets* (4), *weapons secrets* (2), *his former employer's trade secrets*, *Pepsi's trade secrets*, *Kendall-Jackson's trade secrets*, *a designer's trade secrets*, *valuable trade secrets*, *specific trade secrets*, *her teaching secrets*)
- 1 example involves *divulging of trade secrets*

7 examples are passivized:

(142) a. The spit test [. . .] was a trade secret divulged by Rosalind Candlin Benedict
b. The agency declines to say what state secrets have been divulged
c. [. . .] an aloof and distant leader whose secrets have still to be divulged.
d. [. . .] military secrets relating to its current offensive could be divulged by U.N. peacekeepers
e. The secret was probably divulged by Kim Philby
f. These secrets [. . .] were divulged
g. [. . .] prevent trade secrets from being divulged

2 examples are varied in other ways, including an occurrence across a relative clause boundary:

(143) a. There will always be secrets the United States and its allies will never divulge
b. [. . .] the last great city of Western Europe with secrets to divulge

exploit connections

Out of the 22 sentences involving the 'exploiting' of 'connections', only one was 'exploit connections', and one other was 'exploit a connection'. The other 20 examples (91%) all involve variation.

- 10 examples involve other specifiers (such as possessive pronouns and *President Clinton's connection*, *Clinton's connection*, *his mother's connections*, *some of our connections*)
- 4 examples involve modifiers (*political connections*, *eastern European connections*, *real or imagined German connections*)
- 3 examples involve specifiers and modifiers (*his familial connections*, *their political connections*, *any other connection*)

- 2 examples involve specifiers and a compound noun (*his family connections*, *its MCI connection*)

exert influence

Out of 565 sentences involving the 'exerting' of 'influence', 104 are *exert influence*. This corresponds to 18%. It is the most frequent form of this expression (there are only 4 occurrences of *exert the influence* and 12 occurrences of *exert an influence*), so this is not an expression with a canonical form. 52 (9%) of the occurrences are syntactically varied in that they go across clause boundaries and/or are passive, like the examples in (144).

(144) a. [...] the influence that big money exerts on state politics [...]
b. [...] the influence exerted by misleading campaign rhetoric [...]
c. [...] the influence "Superman" comics exerted on my childhood [...]
d. [...] he told of high-pressure sales pitches and subtle persuasion, of influence being exerted from friends, family, teachers.
e. [...] there was no editorial influence exerted by the associations.
f. [...] allegations of improper influence exerted in connection with items pending before the Department of Agriculture [...]
g. [...] if improper political influence was exerted [...]
h. [...] Aleksei, 18, wrote this letter to his family last Dec. 4, appealing [...] that influence be exerted on his behalf [...]
i. [...] the United States will state its views and exert what influence it has to enforce them [...]
j. [...] the influence he expects to exert in the new parliament [...]
k. [...] how much influence does character exert on presidential decisions [...]
l. [...] even specialists on drug policy disagree over how much influence any president can exert [...]

lose status

Out of 272 sentences involving the 'losing' of 'status', only 11 are *lose status*. This corresponds to 4%. The most frequent form is *lose one's status as*.[46] It accounts for 20% of the data. So this is not an expression with a canonical form. In 172 (63%) of the occurrences the noun *status* is modified by adjectives (*diplomatic*, *privileged*, *special*, *legal*, *unique*, *favored*, *tax-exempt*, ...) or is the head of a compound noun (*entitlement status*, *safe haven status*, *most-favored nation status*, *monopoly status*, *investment grade status*, *franchise status*, *superpower status*, ...)
29 examples (11%) are syntactically varied in that they are passive, go across relative clause boundaries, involve coordination, or use *lost* as an adjectival participle. Some examples are given in (145):

(145) a. The Granite State's bellwether status had long ago been lost.
b. [...] giving him a superstar status he never lost [...]
c. [...] the association is struggling to regain the insider status lost early in the Clinton administration [...]
d. [...] Bono can start to rebuild some of the status he lost with his dismal performance in the playoffs last year.
e. [...] President Boris Yeltsin signed a decree Tuesday restoring the legal status of Russian troops in Chechnya which they effectively lost in 1992
f. [...] few foreign firms wish to draw attention to, or lose, their privileged status in China [...]
g. Without certification, Colombia would lose aid and favorable trade status.
h. She then filed a federal suit in July, seeking to regain her lost status.

bring advantages

Out of 62 sentences involving the 'bringing' of 'advantages', only 2 are *bring advantages*. This corresponds to 3%. It is the most frequent form of this expression (there is only one occurrence of *bring an advantage* and one occurrence of *bring the advantages*), so this is not an expression with a canonical form. In 18 examples *advantage(s)*

[46] This assumes that *lose my/your/his/her/its/our/their status as* are all canonical.

is modified by adjectives (*major*, *distinct*, *big*, *enormous*, *other*, *several*, *certain*, *economic*, *great*, *significant commercial*, *short-term political*, *important competitive*, ...) or is the head of a compound noun (*customer service advantages*, *home-court advantage*, *business advantages*, ...). 28 (45%) of the occurrences are syntactically varied in that they go across clause boundaries like the examples in (146):

(146) a. [...] that's an advantage we bring to CBS.
b. Young Louis had all the advantages great wealth can bring.
c. [...] companies seek the advantages that size and economies of scale can bring [...]

end the silence

Out of 64 sentences involving the 'ending' of 'silence', only 2 are *end the silence*. This corresponds to 3%. The most frequent form is *end one's silence*, which accounts for 31% of the data. In 23 examples *silence* is modified by adjectives (*weeklong*, *guilty*, *political*, *public*, *embarrassed*, *remarkable national*, *long, self-destructive*, *official*, *military*, ...) or is the head of a compound noun (*two-day silence*, *20-month silence*, ...).
5 occurrences are syntactically varied in that they are passives or inchoatives, like the examples in (147).

(147) a. The silence was ended, as arranged, with a formation of four fighter jets that thundered over the site once to deliver the nation's salute.
b. The documentary concludes that the long silence is finally ending.
c. This weekend, the silence will end.

Summary

Table 3.3 summarizes the results from the study of non-idioms.

Non-Idioms	**Percent Matched to Canonical Form of Idiom**	**Percent of Most Frequent Form**
reveal the secrets	1%	7%
divulge the information	2%	31%
divulge the secrets	5%	12%
exploit connections	5%	5%
exert influence	18%	18%
lose status	4%	20%
bring advantages	3%	3%
end the silence	3%	31%
Average:	**5%**	16%

Table 3.3: Results from the Study of Non-Idioms

3.9 Summary

In this chapter I have presented data showing that idiom variation is not limited to rare circumstances. On average about 25% of the occurrences of English decomposable V+NP idioms are not in their canonical form, i.e. the idioms show variation that goes beyond mere inflection of their heads. This is a non-negligible amount of variation, so that these data cannot be dismissed as peripheral.

I also document that despite this considerable degree of variation, even decomposable V+NP idioms have a strongly preferred canonical form, which accounts for about 75% of the idioms' occurrences. That contrasts with only about 16% non-varied occurrences among comparable non-idioms. This suggests that speakers know the canonical form of the idiom, and also know which of its properties are essential and which can be varied.

This study also seems to show a clear distinction between the two types of idioms—the difference between the numbers for decomposable and non-decomposable idioms looks quite striking. Whether this difference is statistically significant depends solely on the question of whether the classification into these two types can be considered established independently of these results. Either way, it is clear that there are some highly variable idioms and some less variable idioms, and that both types show a higher degree of canonicity than non-idioms.

Note also that the study of the random sample also shows that 73% of the randomly selected V+NP idioms are semantically decomposable. This confirms the claim in Nunberg et al. (1994:1) that "there are compelling reasons to believe that the majority of phrasal idioms are in fact semantically compositional". But note that these idioms were randomly selected from a list of 750 of the most frequent idioms. It is possible that there are more non-decomposable idioms among less frequent idioms.

Chapter 4

Alternative Approaches to Idioms

In this chapter I discuss the range of possible approaches to idioms within the HPSG framework, especially those that have been proposed in the literature, and also some approaches in other frameworks that do not have a direct analogue in HPSG. The approaches are classified in two ways: whether they represent idioms at the word level or view them as phrasal, and whether the kind of information that gets specified is phonological, syntactic or semantic. If an approach specifies more than one type of information it is considered in the later section. For example, if an approach specifies both phonological and semantic information it is considered in the subsection describing semantic approaches, as they are more powerful.

For each type of approach I start out with what I take to be the most natural way of instantiating it given the general theoretical and formal assumptions outlined in Chapter 1. I then discuss actual instantiations from the literature. Where appropriate, I point out how the approach might be made more powerful, for example by using additional formal tools. I use the idiom *spill the beans* to illustrate each approach, even though some approaches might not be able to describe idioms of this kind adequately.

4.1 Word-Level Approaches

In this section I consider approaches that take lexical entries for idioms to be subtypes of *word*. As mentioned earlier, I call these 'word-level' approaches. The more standard terminology would be 'lexical', but that might lead to confusion since subsorts of *phrase* can be 'lexicalized' as well, in the sense of being entries in the lexicon, or inventory of conventional items.

4.1.1 Multi-Word Lexeme

In this kind of approach a complex expression is treated as a single word syntactically, in this case an intransitive verb:

(148)
$$\begin{bmatrix} \textit{spill\_beans\_verb} \ \& \ \textit{intrans\_verb} \\ \text{PHON} \quad \text{<spill the beans>} \\ \text{SYNSEM} \mid \text{LOCAL} \mid \text{CONT} \mid \text{LZT} \quad \text{<}\textit{i\_spill\_beans\_rel}\text{>} \end{bmatrix}$$

This approach is somewhat similar to that proposed by Chomsky (1980), which also treat idioms as complex verbs, within a transformational framework. There are various problems with such approaches (e.g. McCawley 1981) and they will not be discussed here as they cannot be translated into the HPSG framework. More recently, this kind of approach was proposed within the HPSG framework by Krenn and Erbach 1994, who used it for 'completely fixed' idioms. The example they gave is the German phrase *auf und davon* (literally 'up and away', meaning 'away'):

(149)
$$\begin{bmatrix} \text{PHON} \ \textit{auf und davon} \\ \text{SYNSEM} \mid \text{LOC} \begin{bmatrix} \text{CAT} \begin{bmatrix} \text{HEAD} \ \textit{adj} \ [\text{PRD} \ +] \\ \text{SUBCAT} \left\langle \text{NP}\boxed{1} \right\rangle \end{bmatrix} \\ \text{CONT} \begin{bmatrix} \text{REL} \ \textit{gone-away} \\ \text{GOER} \ \boxed{1} \end{bmatrix} \end{bmatrix} \end{bmatrix}$$

Sag and Wasow (1999:269) used this type of approach for non-decomposable idioms like *kick the bucket*. The example they give is shown in (150).

(150)
$$\left\langle \left\langle \text{kick, the, bucket} \right\rangle, \begin{bmatrix} \textit{iv-lxm} \\ \text{ARG-ST} \left\langle \text{NP}_i \right\rangle \\ \text{SEM} \begin{bmatrix} \text{INDEX } s \\ \text{RESTR} \left\langle \begin{bmatrix} \text{RELN die} \\ \text{CORPSE } i \end{bmatrix} \right\rangle \end{bmatrix} \end{bmatrix} \right\rangle$$

This type of approach might be the best way of dealing with complex items like *by and large* or *ad hoc* that are not compositional or variable, but it is not flexible enough to handle any of the variability that almost all other idioms exhibit. It could not even handle the syntactic insertion of semantically external modifiers into non-decomposable idioms (e.g. *by mere dint of* or *kick the proverbial bucket*). As the data in Chapter 3 showed, it is not the case that *proverbial* is the only adjective that can occur inside such idioms, so that it would not be sufficient to say that *proverbial* is an optional element in the phonology of such idioms.

The only type of variation that can be expressed in such an approach, given the right assumptions about morphology, is inflection of the head. Sag and Wasow say that they "adopt the general convention that morphological functions apply only to the first element of such entries". Inflection in other parts of the idiom cannot be handled unless a special infixation operation is allowed.

Even for invariable expressions like *tit for tat* this approach fails to capture the intuition that the occurrence of *for* in this phrase is related to the lexical item *for*. The same is true for Krenn and Erbach's German example *auf und davon* ('away')—*davon* can mean 'away' by itself, at least in some contexts, but this relationship is not captured in this approach.

Furthermore, unless some kind of wrapping analysis is made available, this approach does not work for idioms with open slots in them, even if the size of the open slot is not variable, as in *give NP the benefit of the doubt*. Further complications arise with idioms like *eat one's heart out*, where the possessive in the open slot has to agree in person and number with the subject of the idiom.

4.1.2 Subcategorizing for the Phonology

In this approach the idiom is represented as a special lexical entry for *spill* that only occurs with the complement *the beans*, where this complement is referred to by its phonology:

(151)
$$\begin{bmatrix} \textit{spill\_beans\_verb} \\ \text{PHON} \quad \text{<spill>} \\ \text{SYNSEM} \mid \text{LOCAL} \quad \begin{bmatrix} \text{CAT} \mid \text{VAL} \mid \text{COMPS} \quad \left\langle \text{NP} \begin{bmatrix} \textit{sign} \\ \text{PHON} \quad \text{<the beans>} \end{bmatrix} \right\rangle \\ \text{CONT} \mid \text{LZT} \quad \textit{<i\_spill\_beans\_rel>} \end{bmatrix} \end{bmatrix}$$

This approach is limited to idioms with invariant complements, and even simple inflection of the complement, such as singular vs. plural, could not be handled without allowing special phonological operations. Variants of this approach have been proposed at various times in the literature, for example by Krenn and Erbach (1994), who used it to describe idioms with 'frozen' complements such as *in Frage kommen* (literally: 'into question come', idiomatically: 'be a possibility'):

(152)
$$\begin{bmatrix} \text{PHON} \ \textit{kommen} \\ \text{SYNSEM} \mid \text{LOC} \begin{bmatrix} \text{CAT} \begin{bmatrix} \text{HEAD} \ \textit{verb} \\ \text{SUBCAT} \left\langle \text{NP}[\textit{nom}]_{\boxed{1}}, \text{PP}\left[\text{PHON} \ \textit{in Frage}\right] \right\rangle \end{bmatrix} \\ \text{CONT} \begin{bmatrix} \text{REL} \ \textit{be-possible} \\ \text{THEME} \ \boxed{1} \end{bmatrix} \end{bmatrix} \end{bmatrix}$$

Approaches of this type have several drawbacks. They require subcategorizing for *sign*s, not *synsem*s, in order to make phonological information available. This fails to preserve the general HPSG assumption of locality in subcategorization.[1] That is, just to accommodate idioms, in this approach all verbal heads are able to impose

[1] It is not clear that my approach or other phrasal approaches make any empirically distinguishable predictions about this matter where idioms are concerned, but at least they do not violate the Lexical Integrity Principle (Bresnan and Mchombo 1995) or the Principle of Phonology-Free Syntax (Zwicky and Pullum 1986). That is, in a phrasal approach the way idioms are treated at least does not affect how ordinary words are treated.

restrictions on the internal constituent structure of their complements, because the PHON and DTRS (DAUGHTERS) are among the attributes of *sign*s and are therefore accessible.

Nothing in this approach ensures that the literal meaning of the string *the beans* is absent from the idiom, as the literal lexical entry matches the phonological information specified. It is possible to specify the idiomatic meaning of *spill the beans* in this special lexical entry for *spill*, but a further mechanism would be needed to discard the semantic contributions of the complement *the beans*. It is not clear how it would be possible to prevent the Semantics Principle from applying, which appends the LZTs of the daughters to form the LZT of the mother, so that the *_bean_rel* of the NP would end up on the LZT of the resulting *head-complement-phrase*. Krenn and Erbach considered the possibility that these subcategorized-for idiomatic phrases did not have any internal structure at all, but they did not discuss how an internal structure could be prevented or how a parser could deal with them in such a case. If the idiom to be described this way is semantically decomposable, there is no way to associate part of the idiomatic meaning with the complement, because the semantics principle cannot apply normally.

Furthermore, this approach does not work for idioms with parts that are optionally modifiable by adjectives or otherwise variable, because such variants would not match the specified phonological string. This is true even of purely syntactic modification, i.e. semantically external modification, like in *kick the proverbial bucket* and *saw major logs*.

As with all word-level approaches depending on subcategorization, it is not possible to represent idioms that involve adjuncts or those that do not have a head in which they could be lexicalized, like *up the creek without a paddle*. None of these approaches can handle occurrences across relative clause boundaries like the example in (153), since the relative pronoun does not have the phonological information required in the subcategorization constraint of the idiomatic verb.

(153) The waves Japanese authorities are making in the currency markets [...]

Note also that among the English idioms I studied there were none with a completely 'frozen' complement that were nevertheless passivizable or otherwise syntactically variable. This may be different in languages like German which have freer word order, but for English there does not seem to be any motivation for an analysis of the type discussed in this section.

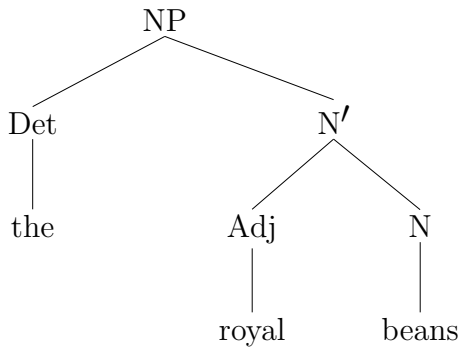
4.1.3 Subcategorizing for the Syntax

This approach is like the previous one in that the idiom is represented as a special lexical entry for *spill*. But *spill* is specified as having *beans* as its complement syntactically and not phonologically. That is, the HEAD-DTR of the NP-complement of *spill* is specified as the lexical entry for the noun *bean*.

(154)
$$\begin{bmatrix} \textit{spill\_beans\_verb} \\ \text{PHON } \langle \text{spill} \rangle \\ \text{SYNSEM} \mid \text{LOC} \begin{bmatrix} \text{CAT} \mid \text{VAL} \mid \text{COMPS} \left\langle \text{NP}[\text{HEAD-DTR } \textit{bean}] \right\rangle \\ \text{CONT} \mid \text{LZT } \langle \textit{i\_spill\_beans\_rel} \rangle \end{bmatrix} \end{bmatrix}$$

This approach suffers from many of the same problems as the previous one, but it can handle inflectional variation in the noun and variation in the specifier, because only the general type for the head daughter is specified. Unless some form of functional uncertainty (Kaplan and Maxwell 1988) is employed, it cannot handle modification by adjectives, because it specifies that *beans* is the immediate HEAD-DTR of the complement. But when *beans* is modified by an adjective, it is instead located one level further down inside the NP, i.e. it is the HEAD-DTR of the HEAD-DTR of the complement of *spill*. That is, an NP like (155) is represented in HPSG as in (156), which does not unify with the constraint on the COMPS list in (154).

(155)



(156)
$$\begin{bmatrix} \textit{head\_spr\_phrase} \\ \text{NON-HEAD-DTRS} \langle\ [\textit{the}]\ \rangle \\ \text{HEAD-DTR} \begin{bmatrix} \textit{head\_adj\_phrase} \\ \text{NON-HEAD-DTRS} \langle\ [\textit{royal}]\ \rangle \\ \text{HEAD-DTR} \begin{bmatrix} \textit{bean} \\ \ldots\text{NUM } \textit{pl} \end{bmatrix} \end{bmatrix} \end{bmatrix}$$

Krenn and Erbach (1994) proposed a variant of this approach that introduces an INDEX feature LEXEME. The example they gave is the German idiom *den Garaus machen* ('finish off').

(157)
$$\begin{bmatrix} \text{PHON } \textit{machen} \\ \text{SYNSEM} \mid \text{LOC} \begin{bmatrix} \text{CAT} \begin{bmatrix} \text{HEAD } \textit{verb} \\ \text{SUBCAT} \left\langle \begin{array}{l} \text{NP}[\textit{nom}]\boxed{1}, \\ \text{NP}[\textit{dat}]\boxed{2}, \\ \text{NP}[\textit{acc,sg}] \begin{bmatrix} \text{DTRS} \begin{bmatrix} \text{HEAD-DTR}\ldots\text{INDEX} \mid \text{LEXEME } \textit{Garaus} \\ \text{MARKER-DTR } \textit{det}\ [\textit{def}] \end{bmatrix} \end{bmatrix} \end{array} \right\rangle \end{bmatrix} \\ \text{CONT} \begin{bmatrix} \text{REL } \textit{kill} \\ \text{KILLER } \boxed{1} \\ \text{KILLED } \boxed{2} \end{bmatrix} \end{bmatrix} \end{bmatrix}$$

In this approach the selection occurs via the value of the feature LEXEME, which avoids the problem of having to know where in the NP the head noun is, since the *index* will be shared and is available at the phrasal level. However, this approach still does not extend to cases where more than just the head noun is part of the idiom but variation is still possible:

(158) a. bark up the wrong tree
b. bark up **another** wrong tree
c. bark up the **proverbial** wrong tree
d. bark up the **same** wrong tree
e. bark up the wrong **political** tree

For such idioms a further feature would be required to allow specifying more than just the lexeme of the head noun in the complement. In this example the adjective *wrong* needs to be specified as well, as it is part of the idiom. A further feature would not be necessary for idioms where the complement is completely fixed and the fixed elements can be found via the DTRS.

This approach still requires subcategorizing for whole signs to allow access to other information via the DTRS, e.g. a fixed PP complement of a noun (*scrape the bottom of the barrel*). Because in this approach the noun *bottom* is identified only as the value of a LEXEME feature, there is no way to state anything about its complements, as would have been possible in the approach in (154). So the LEXEME feature solves some problems, but also creates others. Note also that it is not possible to know how deeply embedded the *barrel* is:

(159) a. Handsome as it is to look at, "The Phantom" conveys the feeling that Hollywood is **scraping the bottom of the comic-strip-hero barrel** here.
b. The Forest Service, meanwhile, is **scraping the bottom of its timber sale barrel** for second-growth timber to substitute for the old-growth here. (from the North American News Corpus)

So even in the approach in (154) it would not be possible to say that the whole complement of *bottom* is fixed.

Furthermore, as in Section 4.1.2, a mechanism is needed for discarding the literal meaning of the string *the beans*. For non-decomposable idioms Krenn and Erbach suggested assigning an empty θ-role to the complement NP, and modifying the Quantifier Inheritance Principle so it ignores items without a thematic role. It might be possible to do something analogous with MRS, i.e. not adding the *rel*s to the LZT when some such marking is absent. However, this would work only when the semantics of the entire NP can be ignored, which is the case for only very few idioms. Even

for non-decomposable idioms like *kick the bucket* one cannot ignore the entire NP, since it may have modifiers added. These cannot be discarded completely even if they get a semantically external interpretation. Furthermore, assigning an empty θ-role is an approach that would not work for semantically decomposable idioms and is not consistent with examples of semantically internal modification.

A different way of avoiding the problem of what to do with the literal meaning of *beans* is to have *spill* instead select for an idiomatic entry for *beans*:

(160)
$$\begin{bmatrix} \textit{spill\_beans\_verb} \\ \text{PHON } \langle \text{spill} \rangle \\ \text{SYNSEM} \mid \text{LOC} \begin{bmatrix} \text{CAT} \mid \text{VAL} \mid \text{COMPS} \left\langle \text{NP}[\text{HEAD-DTR } \textit{i\_bean}] \right\rangle \\ \text{CONT} \mid \text{LZT } \langle \textit{i\_spill\_rel} \rangle \end{bmatrix} \end{bmatrix}$$

However, in that case the independent lexical entry for the idiomatic word *beans* has to be prevented from occurring in the absence of *spill*. A solution to this problem was suggested in Sag and Wasow (1999:266). There, idioms like *keep tabs on* were represented with the lexical entries for *keep* in (161) and for *tabs* in (162).

(161)
$$\left\langle \text{keep} , \begin{bmatrix} \textit{ptv-lxm} \\ \text{ARG-ST} \left\langle \text{NP}_i, \text{NP}[\text{FORM tabs}], \text{PP}\begin{bmatrix} \text{FORM on} \\ \text{P-OBJ NP}_i \end{bmatrix} \right\rangle \\ \text{SEM} \begin{bmatrix} \text{INDEX } s \\ \text{RESTR} \left\langle \begin{bmatrix} \text{RELN} & \text{observe} \\ \text{SIT} & s \\ \text{OBSERVER} & i \\ \text{OBSERVED} & j \end{bmatrix} \right\rangle \end{bmatrix} \end{bmatrix} \right\rangle$$

(162)
$$\left\langle \text{tabs} , \begin{bmatrix} \textit{cn-lxm} \\ \text{SYN} \begin{bmatrix} \text{HEAD} \begin{bmatrix} \text{FORM tabs} \\ \text{AGR} \begin{bmatrix} \text{PER 3rd} \\ \text{NUM pl} \end{bmatrix} \end{bmatrix} \\ \text{SPR} < > \end{bmatrix} \\ \text{ARG-ST} < > \\ \text{SEM} \begin{bmatrix} \text{MODE none} \\ \text{INDEX none} \\ \text{RESTR} < > \end{bmatrix} \end{bmatrix} \right\rangle$$

In this approach the lexical entries for idiomatic words like *tabs* and *beans* have no meaning associated with them, and in particular their INDEX value is *none*, which is intended to make them ineligible as complements to ordinary verbs. However, as the authors say, this analysis was simplified for textbook purposes and does not capture their intuitions (expressed in Nunberg et al. 1994) that in these semantically decomposable idioms parts of the idiom carry parts of the meaning and can be internally modified.

This type of approach also cannot handle occurrences across relative clause boundaries where the idiomatic word is a complement of an ordinary verb, and the idiomatic verb selecting for it is only present in a subordinate relative clause.

4.1.4 LFG

The approach in Bresnan (1982:46) is very similar to the one just discussed:

(163) keep: V, 'KEEP-TABS-ON ((SUBJ),(OBJ))'
(OBJ FORM) $=_c$ TABS

In both cases the idiomatic noun is identified via a special FORM feature. In the HPSG version it is a HEAD feature and therefore available at the phrasal level, so that the description is consistent with NPs that have inserted adjectives. In the LFG version this is expressed as a constraint equation, which can also deal with syntactically inserted adjectives. But this only helps with cases of semantically external

modification. In both approaches, words bear the FORM feature only if they lack intrinsic meaning, in order to guarantee that idiom chunks appear only with idioms. So this LFG approach is also unable to deal with semantically decomposable idioms. If the LFG feature PRED were used instead of FORM, this would not be a problem. However, in that case another mechanism would have to ensure that *tabs* cannot occur by itself. It would also be necessary to develop a mechanism to 'throw away' the literal meanings, and to do so in such a way that modifiers and unusual specifiers survive this procedure and correctly modify the idiomatic meaning. To my knowledge noone has worked out and published the details of such an analysis, in LFG or HPSG.

Bresnan (2001) assumes a similar word-level approach to idioms, except that the constraints are stated in terms of argument structure. The examples given are for Dyirbal, an ergative language, so the patient is actually the subject.

(164) **bana-l** ⟨ *ag* *pt* ⟩
'break'
|
buŋgu
'knee'

buŋgu + **bana-l** 'bend over, fold'

Bresnan (2001) is a textbook, so this analysis may well be simplified. Furthermore, idioms are discussed only as evidence for a relational design of grammar, and no further details about handling idioms are given. Relative clauses are also not discussed. For these reasons it is hard to evaluate what exactly the problems with this approach may be.

Her et al. (1994) discuss another LFG approach to idioms developed for the purposes of machine translation. In this approach sentences are first parsed as usual, and then undergo special treatment during transfer.

> As far as the syntax, or parsing, of the source language is concerned, there is no need to treat transfer idioms as in any way special. (Her et al. 1994:209)

In order to describe whether an idiom is syntactically variable, certain types of information have to be specified in the f-structure, since in their translation system only the source f-structure enters the transfer stage. The example f-structure for *kick the bucket* is shown in (165). Presumably the features SPFORM and ADJS (ADJUNCTS) are

used only for this purpose and are not needed elsewhere in the grammar. Note that there is no idiomatic meaning associated with this f-structure, because it is designed to match a literal parse. Instead, a transfer rule maps this f-structure onto Chinese *si3* 'to die'.

(165)
$$\begin{bmatrix} \text{FORM} & \text{´kick´} \\ \text{PRED} & \text{<SUBJ OBJ>} \\ \text{VOICE} & \text{ACTIVE} \\ \text{OBJ} & \begin{bmatrix} \text{SPFORM} & \text{´the´} \\ \text{FORM} & \text{´bucket´} \\ \text{NUMBER} & \text{SG} \\ \text{ADJS} & \text{NONE} \end{bmatrix} \end{bmatrix}$$

This is obviously a computational and not a theoretical approach, since from a linguistic point of view it does not make sense to say that idioms have to be dealt with only in the context of translation: *kick the bucket* can have the idiomatic meaning 'die' even to a monolingual speaker of English. An approach of this type could be extended to cover idioms language-internally by adding a system similar to the transfer rules that 'translates' *kick the bucket* to *die* after parsing the sentence using only literal lexical entries. But this works only for idioms that have a literal parse, and it is psycholinguistically implausible because it would predict that idioms are understood more slowly than non-idioms since extra work has to be done applying the special translation rules.

4.1.5 Subcategorizing for the Semantics

In this approach the idiom is also represented as a special lexical entry for *spill*, but what is specified about the complement is its main semantic contribution, using the MRS feature KEY. In an NP this always points to the meaning of the head noun, however deeply embedded it is.

(166)
$$\begin{bmatrix} \textit{spill\_beans\_verb} \\ \textsc{phon} \ <\text{spill}> \\ \textsc{synsem}\,|\,\textsc{loc} \begin{bmatrix} \textsc{cat}\,|\,\textsc{val}\,|\,\textsc{comps} \left\langle \text{NP}[\textsc{loc}\,|\,\textsc{cont}\,|\,\textsc{key}\ \textit{i\_bean\_rel}] \right\rangle \\ \textsc{cont}\,|\,\textsc{lzt}\ <\textit{i\_spill\_rel}> \end{bmatrix} \end{bmatrix}$$

This approach is less general than the previous one because selecting for the meaning of the idiom part presupposes that that meaning exists as a separate entity. Non-decomposable idioms like *kick the bucket* have an unanalyzed *i_kick_bucket_rel* corresponding roughly to 'die' in their semantics, which cannot be distributed over their syntactic parts. But this approach requires an idiomatic *i_bucket_rel*, so it is not appropriate for non-decomposable idioms.

Furthermore, word-level approaches that presuppose lexical entries for idiomatic words need an additional mechanism to make sure that these idiom parts cannot occur by themselves. Without such a mechanism, sentences like (167) are predicted to occur with the idiomatic meaning *i_beans_rel*, i.e. something like 'secrets', for *beans*.

(167) #I am very good at keeping beans.

It is not possible to say that the LZTs of valence elements of non-idiomatic verbs are not allowed to contain *i_rel*s, because that would rule out examples like (168), where the idiomatic word *backseat* is a complement of the non-idiomatic word *mourn*:

(168) a. He mourns the backseat that books have taken to movies [...]

One might wonder why one could not assume that the lexical entry for *beans* has an additional feature I-CONTENT for the idiomatic meaning, and that *spill* subcategorizes for a complement with that I-CONTENT. However, there are several problems with such an approach. First, it would be necessary to complicate the semantics considerably to ensure that in the idiomatic cases, the literal CONTENT is ignored and the idiomatic content is used instead. Secondly, some words have more than one idiomatic meaning, e.g. *strings* can mean something like *connections* in *pull strings*, something like *control* in *pull the strings*, something like *conditions* in *no strings attached*, and further meanings in *tied to one's mother's apron strings* and *hold the*

purse strings.[2] So one would need to further complicate this approach to account for these cases. This might be done by having multiple lexical entries for *strings*, each corresponding to only one idiomatic meaning. But then one would have to avoid multiple parses with the literal meaning in some way. Or it might be done by having multiple I-CONTENT features: I-CONT-1, I-CONT-2, I-CONT-3, I-CONT-4, I-CONT-5, etc., but that would require further complicating the semantics to ensure that the correct meaning is selected. In addition, many of those features would end up being unused for nouns with fewer idiomatic meanings. I have not tried to work out these alternatives in detail as they do not sound appealing to me.

One might think that the approach in (166), as well as the one in Sag and Wasow (1999) which uses a HEAD feature FORM, at least respects the locality principle. These approaches do not require subcategorizing for whole signs because the HEAD and KEY features, respectively, are available at the level of the entire NP.

However, for many idioms it is necessary to specify more than just the head noun of a complement. Sometimes specifiers, modifiers, and even complements of the noun have to be fixed too, (e.g. *scrape the bottom of the barrel*). It is not the case that when one of these is fixed, the whole complement is 'frozen'. There are cases where more than just the head noun is fixed, but variation is still possible, e.g. *bark up the same wrong tree*, which makes locating the modifier *wrong* within the DTRs impossible without additional formal tools.

One possible way to describe these other fixed items in this kind of approach is to invent new features pointing to them. The LinGO grammar already uses a feature COMPKEY which always points to the head noun of the NP complement, so that (169) amounts to the same thing as (170).

(169)
$$\begin{bmatrix} \textit{spill\_beans\_verb} \\ \text{PHON} \quad \langle \text{spill} \rangle \\ \text{SYNSEM} \mid \text{LOCAL} \mid \text{CONT} \quad [\text{COMPKEY } \textit{i\_bean\_rel}] \end{bmatrix}$$

[2]This is by no means rare. Looking at the index of idioms dictionaries one can see that many words occur in more than one idiom. For example, *sleeve* occurs in *have something up one's sleeve*, *roll up one's sleeves*, *wear one's heart on one's sleeve* etc., and *straw* occurs in *draw the short straw*, *clutch at straws*, *the straw that breaks the camel's back*, and others.

(170)
$$\begin{bmatrix} \textit{spill\_beans\_verb} \\ \text{PHON} \quad \text{<spill>} \\ \text{SYNSEM} \mid \text{LOCAL} \mid \text{CAT} \mid \text{VAL} \mid \text{COMPS} \quad \left\langle \text{NP}[\text{LOCAL} \mid \text{CONT} \mid \text{KEY} \ \textit{i\_bean\_rel}] \right\rangle \end{bmatrix}$$

Some further features like OCOMPKEY for an oblique complement's main semantic relation can be added without problems. Since idioms sometimes specify their subjects (e.g. *a little birdie told me*), one would need to introduce a SUBJKEY feature as well.

But as we saw in Section 2.2.9, adjectives and specifiers sometimes need to be fixed too. So additional COMPADJKEY and SUBJADJKEY features are required, plus COMPSPRKEY and SUBJSPRKEY features for specifiers.[3] It is not so clear that it would be possible to ensure that these always point to the right items, given that specifiers can be complex and noun phrases can contain multiple modifiers. The fixed modifier may not have to be in a particular position: both *bark up the* ***proverbial*** *wrong tree* and *bark up the wrong* ***political*** *tree* are attested.
Furthermore, there are idioms in which the PP complement of the noun is fixed:

(171) a. scrape the bottom of the barrel
b. see the color of his money

These examples would require the additional features COMPCOMPKEY[4] (to select for the preposition *of*), COMPCOMPCOMPKEY (to select for the head noun *barrel* of the NP within that PP), and COMPCOMPCOMPSPRKEY (to select for the specifier *the* of that noun). Further features would be required for (172), and it probably does not stop there.

(172) tied to one's mother's apron strings

One might think that this proliferation of features could be avoided if functional uncertainty or a set membership operation were available. Such an operation would make it possible to just require a particular *rel* to occur somewhere on the LZT. But

[3] These features would correspond roughly to the SPFORM and ADJS features in the LFG approach in Her et al. (1994).

[4] It is not possible to avoid this by going down one level explicitly, as in (170), and then using the COMPKEY feature, since the COMPS list of the NP *the bottom of the barrel* is empty because the valence requirement has been satisfied.

that does not seem sufficient, since presence of the relevant items somewhere in the complement is not sufficient. For example, (173) is not an instance of the idiom *scrape the bottom of the barrel*, even though the *i_barrel_rel* does occur on the NP's LZT.

(173) #She scraped the bottom of the bowl that was located in the wooden barrel.

This could perhaps be prevented by specifying what semantic relation these *rel*s have to stand in, as in the constructional approach to be developed in Chapter 5. But the word-level approach still is not powerful enough to handle examples that do not involve a head, e.g. *butterflies in one's stomach*. As was shown in Chapter 3, there is no lexical entry at the word level where the relevant relationship could be stated. A word-level approach cannot handle occurrences across relative clause boundaries either, since the relative pronoun would not meet the subcategorization specification—the INDEX is shared between it and the modified noun, but the *rel* is not, either in the approach in Sag (1997) or in Pollard and Sag (1994). And while it can handle raising constructions, in which the whole CONTENT is shared, it cannot handle idioms occurring in control (equi) constructions, in which only the INDEX is shared.[5] Word-level approaches also require that all adverbs and adjuncts are made available on the complements list because such items can be obligatory parts of idioms.

4.1.6 GPSG - Partial Functions

In GPSG (Gazdar et al. 1985:238) the semantic translation of *spill* is a partial function defined only at one argument, the idiomatic sense of *the beans*. The idiomatic sense of *the beans* is prevented from occurring with other verbs by giving ordinary verbs a semantics that is also a partial function not defined for such idiomatic complements. However, this is not consistent with the existence of examples like (174).

(174) [...] Robert McNamara's new book **justified all the strings Clinton pulled** as a young man [...] (from the North American News Corpus)

[5]It would be possible to handle these examples if there were special index values for each idiomatic word, and these were used for selection instead of the *rel*s. However, this would result in a very rich set of *index* types which are quite unlike the kinds of selectional restrictions (animate, human, etc.) that have been encoded this way in the past, and that would exactly duplicate the set of idiomatic *rel* types.

In this example *the strings* is the complement of *justified*. The reason that this is possible is that the idiomatic occurrence of *the strings* is 'licensed' by the verb *pulled*, which occurs in the relative clause and stands in the right semantic relationship to *the strings*. But this is not handled by the GPSG approach.

A related problem with the GPSG approach is that it does not allow for modification of idiomatic complements like *the hatchet* by relative clauses, as in (175):

(175) The Robinsons have buried the family hatchet that made the band's 1994 album, "Amorica," such an angry and depressing work

If the semantic function of literal verbs like *make* has to be allowed to take idiomatic arguments like *hatchet* to account for such examples, then it is not clear how to prevent it from doing so in the absence of *bury*, as in (176):

(176) #The family hatchet made the band's album a depressing work.

It is also unclear how exactly this approach allows for modification by adjectives. That is, if *spill* is a partial function defined only at one argument, the idiomatic sense of *the beans*, then it is not clear how *the royal beans* can be allowed. Pulman (1993:258) assumes that GPSG allows only those combinations of functions and arguments that "have all the literal or all the idiomatic senses". Note that it is not sufficient to require that *the beans* is present somewhere in the argument of *spill*, as (177) does not have an idiomatic interpretation.

(177) #She spilled the soup that contained the beans.

4.1.7 Summary of Problems with Word-Level Approaches

Almost all word-level approaches that represent idioms as special lexical entries for the head verbs and specify particular facts about their complements have problems with almost all of the data discussed in the section 'Need for Phrasal Pattern' in Chapter 2.

One problem with all these approaches is that only one of the words in the idiom is constrained to co-occur with the others—and there is no satisfactory mechanism for controlling the distribution of the others. For example, if *pull strings* is analyzed as a special idiomatic lexical entry for *pull* that only takes the idiomatic word *strings* as

its complement, then there needs to be a mechanism that prevents idiomatic *strings* from occurring without *pull*. Note also that these approaches require the selecting word to be made special even when it actually has the literal meaning, like *miss* in *miss the boat*.

Because these approaches use the usual subcategorization mechanisms by which verbs select their complements, idioms pose problems in that they require specifying information that is never necessary for ordinary subcategorization, in the form of adverbs and adjuncts, as well as items included at various levels of embedding in the complements, such as adjectives and specifiers.

Another problem is restricting the flexibility and specifying the canonical form of idioms. Such approaches would need to develop special mechanisms to be able to express all the ways in which idioms and canonical forms of idioms can be fixed. In the absence of such a mechanism, these approaches predict that idioms are just as flexible syntactically as non-idioms because they are treated in essentially the same way. These approaches have nothing to say about the psycholinguistic evidence that idioms are processed faster than non-idioms, and that canonical forms of idioms are processed even faster.

Word-level approaches also cannot express the metaphorical mapping between the literal and figurative meanings of the idiom as a whole. If such a relationship is to be established at all in these approaches, the mapping has to be expressed at the level of the individual words. That is, the literal lexical entry for *beans* with its 'legume' meaning could be related to the idiomatic lexical entry for *beans* with its 'secrets' meaning, but this cannot be done at the level of the whole idiom, so the fact that this mapping only exists as part of the whole idiom is not expressed, and there is no mapping for the whole metaphor.

Word-level approaches also have problems with some types of variation, in particular occurrences across relative clause boundaries, and variations like *let the cat out of the bag - the cat is out of the bag*, that are best accounted for by treating the idiom as syntactically headless.

The reason that such approaches have been predominant in spite of all the problems associated with them is that it was thought to be impossible to deal with the

variation data in a phrasal approach. But Nunberg et al. (1994) showed only that the relationship between the parts of most idioms is semantic in nature, not that this relationship has to be expressed by giving these parts independent lexical entries. That may have seemed to be the only possible analysis at the time, in the absence of a semantic formalism like MRS which allows specifying the relationship between a verb and the head of its complement in a non-configurational way consistent with further modification and syntactic variations (see Chapter 5).

4.2 Phrasal Approaches

We have seen that no word-level approach can adequately handle the data, and that there are various other problems associated with such approaches. So a phrasal approach of some sort is needed, and this section examines how phrasal approaches fare in comparison.

4.2.1 (Partially) Fixed Phonology

In this approach the phonology of the entire VP is specified as fixed. To my knowledge an approach of this type has not been proposed in the formal literature, so it will not be discussed in much detail.

(178)
$$\begin{bmatrix} \textit{spill\_beans\_phrase} \\ \textsc{phon} \quad \langle \text{spill the beans} \rangle \\ \textsc{synsem}\,|\,\textsc{local}\,|\,\textsc{cont}\,|\,\textsc{lzt} \quad \langle \textit{i\_spill\_rel},\ \textit{i\_bean\_rel} \rangle \end{bmatrix}$$

As in the multi-word lexeme approach, it may be possible to handle inflection of the head verb here, but further variation would require allowing operations on PHON values of a kind yet to be developed. In particular, one would need a mechanism to allow for the inflection of certain parts of the phonology, and to insert adjectives and specifiers in some places and not others. It is unclear how these constraints could be described in a purely phonological way. And such a description could probably not be made consistent with passivized examples or idioms occurring across relative clause boundaries.

In addition, this approach would be unable to deal with idioms like those in (179) that are not complete constituents.

(179) a. sweep NP under the carpet
b. play cat and mouse with NP
c. fly in the face of NP
d. breathe down NP's neck
e. blow one's cool
f. NP's days are numbered

It is hard to see how the open slots here could be characterized in a purely phonological way, when their locations are clearly syntactic in nature.

In addition, this approach does not address the problem of how to avoid including the literal meanings of the words, as a phrase constructed from the literal lexical entries would match the phonological information specified. This is the most obvious difference between this approach and the word-level approach of specifying 'multi-word lexemes', where literal meanings are not a problem, because a VP consisting of literal words assembled by the parser does not unify with a word-level description.

4.2.2 (Partially) Fixed Syntax

In this approach the idiom is represented as a VP in which idiomatic *spill* is the HEAD-DTR and idiomatic *beans* is the HEAD-DTR of the COMP-DTR. To my knowledge an approach of this type has not been proposed in the formal literature. Some transformational approaches might be said to fall in this category, but they have somewhat different properties because, e.g., an idiom can be passivized even if it is represented phrasally.

(180) $\begin{bmatrix} \textit{spill\_beans\_phrase} \\ \text{HEAD-DTR } \textit{i\_spill} \\ \text{COMP-DTRS } \left\langle [\text{HEAD-DTR } \textit{i\_bean}] \right\rangle \end{bmatrix}$

Note that number is underspecified for the complement, and so is the specifier of *bean*. The only types of variation this approach can handle are inflection, variation

in the specifier, and open slots of non-variable size. It works only for idioms that do not passivize, because in a passive sentence the *beans* would not be among the COMP-DTRS. It also does not allow modification by adjectives, because in that case the HEAD-DTR of the COMP-DTRS would not be the head noun *i_bean*, but the N′ that includes the adjective, as we saw in the case of the corresponding word-level approach.

An approach that selects for *rel*s instead of top-level types is just a variant of this approach, and suffers from the same problems. Introducing an *index* feature LEXEME would lead to an approach with fewer problems than at the word level, since no locality violation would be necessary because in this approach things are specified via the COMP-DTRS, not the COMPS list.

(181)
$$\begin{bmatrix} \textit{spill\_beans\_phrase} \\ \textsc{head-dtr}\ \textit{i\_spill} \\ \textsc{comp-dtrs}\ \left\langle [\textsc{head-dtr}\ldots\textsc{head}\ldots\textsc{lexeme}\ \textit{bean}] \right\rangle \end{bmatrix}$$

But this still does not extend to cases where more than just the head noun is fixed and variation is still possible. In order to handle such data, it would be necessary to introduce functional uncertainty, to be able to express the fact that an element has to occur somewhere in the daughters. Alternatively, a set-membership operation could be used to allow for a *rel* to occur somewhere on the LZT, essentially achieving the same effect. But as with the word-level approaches, 'somewhere' on the LZT or in the daughters is not good enough—examples like (182) do not have an idiomatic interpretation, even though *red* is on the LZT and among the daughters of the complement of *roll out*.

(182) #He rolled out the carpet with the red stains.

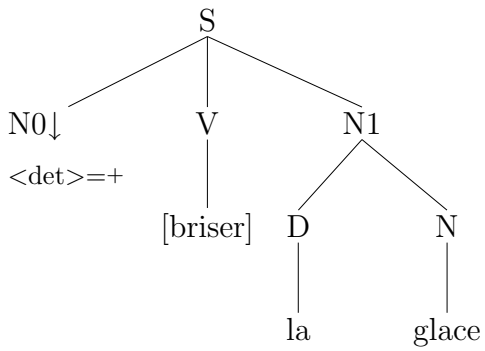
So additional constraints would need to be added to make sure that *red* and *carpet* stand in the right relationship to each other. Note that once set-membership is available and additional semantic relationships are specified, such an approach would be of the 'fixed semantics' type, and one can just as well look at the LZT of the whole phrase (see Section 4.2.4). It should be noted that the constructional approach I am proposing is not inconsistent with syntactic specification. In fact, for idioms that are

not syntactically flexible, both syntactic and semantic information may need to be specified.

4.2.3 TAG

In the Lexicalized Tree Adjoining Grammar (TAG) approach described in Abeillé (1995), idioms correspond to initial elementary trees similar those used for non-idiomatic verbs, except that the frozen complements are included as well. An example is the elementary tree for the idiom *briser la glace* 'break the ice' given in (183).

(183)



These elementary trees are combined by substitution or adjunction. Substitution inserts an initial tree at a leaf node. Adjunction inserts an auxiliary tree at any node labeled with the right category. Because specifiers, modifiers, and raising verbs are treated as auxiliary trees, such modifications of idioms can be handled by the TAG approach. There are also lexical rules that generate further elementary trees, for example for passive constructions and relative clauses, so these types of variation can be handled by the TAG approach as well. Abeillé did not mention handling inflection of the complement, but the TAG approach can handle it because the terminal nodes like [glace] are lemmas and therefore consistent with various inflected forms.

So this approach is good at handling variation. This is not surprising since verbal idioms are handled in essentially the same way as ordinary verbs, so that all the mechanisms available in the grammar for such verbs are available for idioms too. However, there is a flip side to this: even though the approach is 'phrasal', it has many of the problems usually associated with word-level approaches, precisely because the grammar in this approach is designed to work with phrasal elements.

For example, it is not clear how to restrict the flexibility of idioms when this is necessary, and Abeillé did not discuss this. Feature structures can be associated with each node in an elementary tree, and constraints on combining trees can be expressed in terms of success or failure of unification of these features. So it might be possible to prevent adjunctions by inventing new features for each type of adjunction and specifying the idioms correspondingly, to match only the adjunctions allowed. Note that this would complicate the whole grammar, not just the grammar of idioms, because all the auxiliary trees would have to be marked this way.

The next question is how to prevent lexical rules from applying. Again it might be possible to invent a feature for each lexical rule, and use it to mark idioms which do not undergo that rule as ineligible for it. The same problems arise when trying to describe the canonical forms of idioms.

Idioms that contain adjuncts violate the definition of elementary tree, which says that they must correspond to complete argument structures. These idioms cannot be handled in the TAG approach without changing that definition, and this would have consequences for the rest of the grammar.

Idioms that do not contain a syntactic head are even more of a problem. Elementary trees are defined as having at least one lexical anchor, and as corresponding to one non-vacuous semantic unit. These assumptions are probably too central to the TAG approach to be changed.

The approach also does not establish a link between the literal and the idiomatic lexical entries, and therefore does not predict that irregular inflections and many other properties of words occurring in idioms are the same as those of the corresponding literal words. For the same reason it does not establish a metaphorical mapping between the literal and figurative meaning of the idiom. It also seems to be impossible to represent idiom families in such a way that semantic relationships among the literal words are exploited. The only way to specify idiom families in this approach seems to be the one Abeillé actually uses: noting them in the lexicon by means of disjunction over lexical heads.

4.2.4 (Partially) Fixed Semantics

In this approach, which is like one of the approaches discussed in Copestake (1994), only the semantic relationship between the parts of the idiom is specified.

(184)
$$\left[\begin{array}{l} \textit{spill\_beans\_phrase} \\ \\ \textsc{synsem}\,|\,\textsc{loc}\,|\,\textsc{cont}\,|\,\textsc{lzt}\left\langle \ldots, \left[\begin{array}{l}\textit{i\_spill\_rel}\\ \textsc{und}\ \boxed{1}\end{array}\right], \left[\begin{array}{l}\textit{i\_bean\_rel}\\ \textsc{inst}\ \boxed{1}\end{array}\right], \ldots \right\rangle \end{array}\right]$$

The approach proposed in this dissertation is of this general kind. It is described in detail in Chapter 5. But there are several differences that enable it to handle non-decomposable idioms and not to require additional lexical entries for parts of idioms, thereby avoiding the problem of how to ensure that these parts do not occur outside the idiom (see Section 5.4).

4.2.5 Jackendoff

The approach to idioms in Chapter 7 of Jackendoff (1997) is similar in spirit to Copestake (1994), Riehemann (1997), and the approach proposed here. But Jackendoff's approach cannot be translated directly into HPSG because some of the underlying assumptions are different, and not enough detail is given to see how it is supposed to work.

Jackendoff starts from the observation that there are large numbers of idioms and other fixed expressions that speakers know, and that must be represented in some way. Projecting from a sample of 600 puzzle solutions from the U.S. TV show *Wheel of Fortune*, he estimated that in the 10 years this show had been on the air it used ten to fifteen thousand puzzles "with little if any repetition, and no sense of strain—no worry that the show is going to run out of puzzles, or even that the puzzles are becoming especially obscure". These puzzles include compounds, names, clichés, titles, and quotations in addition to idioms. In addition, speakers know the lyrics of many songs, advertising slogans, and many adjective-noun collocations like *heavy/*weighty smoker*, and *weighty/*heavy argument*.

Similar arguments could probably be made about crossword puzzles, which also frequently require knowledge of these types of fixed expressions. In crossword puzzles

some short words with many vowels like *Oreo* are frequently used. But apart from those, crossword puzzle makers can also rely on the solvers' knowledge of a seemingly never-ending supply of fixed expressions. For example, the crossword puzzle clue *in good* ___ , limits the set of possible answers to a manageable number, including *in good faith*, *in good time*, and *in good stead*, and crucially excludes many words that can freely occur in such a PP, such as *in good schools*, *in good seasons*. To add another observation to this, the majority of the 8500 idiomatic expressions in NTC's American Idioms Dictionary are familiar even to me as a non-native speaker, and most native speakers probably know a very large percentage of them.

Jackendoff argued that these fixed expressions need to be seen as part of the linguistic lexicon in order to avoid duplicating linguistic knowledge in another module:

> First it is worth asking why the "received view" would want to exclude clichés, quotations, and so forth from the lexicon. The likely reason is for the purpose of constraining knowledge of language—to make language modular and autonomous with respect to general-purpose knowledge. But in fact, the boundaries of modules must be empirically determined. One can't just "choose the strongest hypothesis because it's the most falsifiable" and then end up excluding phenomena because they're not "core grammar" (i.e. whatever one's theory can't handle). In order to draw a boundary properly, it is necessary to characterize phenomena on *both* sides of it, treating phenomena "outside of language" as more than a heterogeneous garbage can.
>
> In the present case, in order to draw a boundary between the theory of words and that of fixed expressions, it is necessary to show what the theory of fixed expressions is like and how it is distinctively different from the theory of words. In fact, the theory of fixed expressions must draw heavily on the theories of phonology, syntax, and semantics in just the way lexical theory does, and it must account for a body of material of roughly the same size as the word lexicon. Hence significant generality is missed if there is such duplication among theories—and little motivated constraint is lost by combining them into a unified theory. In the course of this Chapter, we will see that most of the properties of fixed expressions (and idioms in particular) are not that different from properties found in semiproductive derivational morphology. (Jackendoff 1997:157)

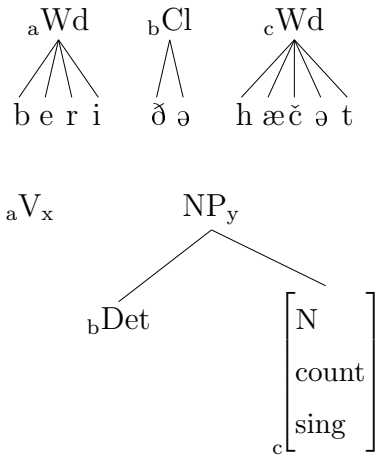
Jackendoff then argued that a word-level account of idioms does not work:

> A possibility [...] is to more or less simulate the listing of *kick the bucket* with monomorphemic lexical entries, by stipulating that *bucket* has a special interpretation in the context of *kick* and vice versa [...] Given a body of fixed expressions as numerous as the single words, such clumsy encoding should be suspect. Indeed, because of the complications of such "contextual specification," no one (to my knowledge) has really stated the details. In particular, it is (to me at least) totally unclear how to "contextually specify" idioms of more than two morphemes. For instance, in order to specify *let the cat out of the bag* one word at a time, contextual specifications must be provided for *let*, *cat*, *out of*, and *bag*, each of which mentions the others (and what about *the*?). Moreover, the correct configuration must also be specified so that the special interpretation does not apply in, say *let the bag out of the cat*. I conclude that this alternative rapidly collapses of its own weight. (Jackendoff 1997:160)

He concluded that a constructional approach was necessary, and that there might not be a clear boundary between lexicon and grammar. Like Chapter 6 of this dissertation, Jackendoff argued in his Section 7.7 that idioms can interact with constructions in interesting ways, and he looked at idioms like *cry one's eyes out* and *scare the daylights out of someone* that are specializations of the resultative construction.

Unfortunately Jackendoff did not work out the details of this constructional approach. The representation he gave for the idiom *bury the hatchet* is shown in (185):

(185)



[RECONCILE ([ ]$_A$, [DISAGREEMENT]$_y$)]$_x$

In this approach the phonological, syntactic, and semantic information about the idiom are all given separately, but the parts are linked to each other as indicated by the subscripts. The phonological and syntactic parts of the representation do not establish any relationship between *bury* and *the hatchet*. Instead this relationship is expressed only in the semantic representation. This semantic representation is the same for passivized sentences, so that this representation for the idiom can handle passive as well as active examples.

However, it is unclear whether it can handle other types of variation, because it is not clear exactly how this representation interacts with the rest of the grammar. According to Jackendoff "a lexical entry enters a grammatical derivation by being unified simultaneously with independently generated phonological, syntactic, and semantic structures" (Jackendoff 1997:161). In the context of idioms with open slots such as *take someone to the cleaners* he expanded a bit more on how this unification would work: "Suppose we assume that unification preserves sisterhood and linear order but not adjacency." (Jackendoff 1997:161)

But it is unclear how this should be formalized, and it would not handle much of the variability data. For example, it could not handle inserted adjectives, even ones like *proverbial*, because *the* and *hatchet* would no longer be sisters. On page 169 Jackendoff says that 'proverbial' does not interfere with unification in syntax and can be semantically integrated with the LCS. But he gave no indication of how this could be achieved. He did not claim that this would work for adjectives that modify *hatchet* semantically, as e.g. in *bury the legal hatchet*, which means 'reconcile the legal disagreement' (or 'end the legal conflict', whichever paraphrase one prefers) and not 'legally reconcile the disagreement', 'lawfully reconcile the disagreement', or 'in the domain of (the) law, reconcile the disagreement'. It is likely that in his semantics the fact that the adjective *legal* modifies *hatchet* would be represented by embedding, i.e. (LEGAL(DISAGREEMENT)), which would not unify with the semantics given for the idiom.

Saying that *the* is obligatorily part of the idiom would exclude the 19 corpus examples without a specifier or with other specifiers, and saying that *hatchet* must be singular excludes another 7 corpus examples. So of the 35 non-canonical occurrences

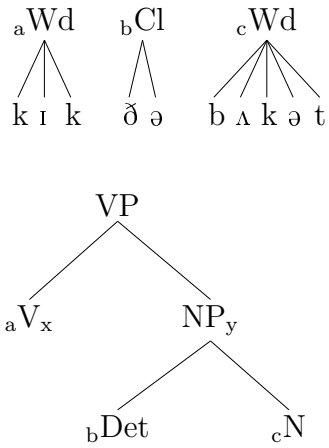
of this idiom observed in the corpus, this approach can handle only the 7 passivized ones. Overall, it fails to account for 19% of the occurrences of this idiom in the corpus. (For more detail on the corpus data for this idiom see Chapter 3.)

Note also that Jackendoff says that only the phonological information and its mapping to the syntactic information is related to that of the literal words (p. 162). This has several disadvantages. Because Jackendoff did not establish a relationship between the idioms and the whole lexical entries of the literal words, he could not describe idiom families, and he did not express the metaphorical mapping at all. The phonological relationship that is expressed is not much more than stipulation. For an idiomatic word related to a word with multiple literal meanings, it would not matter which one the phonological information is taken from. For example, *bank* in *laugh all the way to the bank* could just as well be related to the phonology of the word *bank* meaning 'rising ground bordering a river' as to the phonology of the word *bank* meaning 'financial institution'.

If only the phonological information is shared, then it would be accidental that the syntactic information is usually the same, for example the fact that idiomatic *bury* is a verb and idiomatic *hatchet* is a noun. So one needs a relationship between the idiom and the whole literal lexical entry, but that relationship is not simple identity. It also needs to convey which properties are different. Otherwise this approach would not work for idioms where the syntactic information for the idiomatic word is different from that for the literal word, e.g., where a noun is a mass noun instead of a count noun. Jackendoff seemed to agree with this elsewhere (Jackendoff (1997:166)) when he said that overrides were necessary for compounds and for idioms. But he did not develop his theory of idioms in this way.

This approach also cannot handle idioms that are somewhat but not totally inflexible, which is the case for most non-decomposable idioms. For idioms like *kick the bucket* Jackendoff gave the representation in (186):

(186)



[DIE ([ ]$_A$)]$_X$

This approach cannot handle any variation of this idiom, including semantically external modification like *kick the proverbial bucket*.

It is not just Jackendoff's approach to idioms that is similar in spirit to the approach proposed in this dissertation. He also saw phrasal fixed expressions as related to derivational morphology. I believe that the approach developed in this dissertation captures many of Jackendoff's intuitions about morphology and idioms, but using the formal tools of HPSG and MRS these can be given a somewhat different and more precise formulation.

4.3 Pulman - Quasi-Inference

Pulman (1993) developed an interesting analysis that falls under neither of the two dimensions used for classifying these approaches. In principle, it is compatible with HPSG, although it would not leave the grammar much work to do and would require additional machinery.

In Pulman's system, the first step is to parse idioms and assign them their literal compositional semantics. If this results in a logical form that entails the antecedent of one of the idiom rules, that rule can be applied. An example of such an idiom rule is given in (187):[6]

[6]The symbol ≈ is supposed to have an interpretation similar to an implicational → except that

(187) $\forall x, y\ [cat(x) \wedge bag(y) \wedge out\text{-}of(x, y)] \approx \exists a, z\ [secret(z) \wedge revealed(a, z)]$

This approach can deal with a broad range of variation data, including raising and occurrence across relative clauses. It does not require separate lexical entries for the idiomatic senses of words.

Unfortunately the approach overgenerates significantly. One problem is that other lexical items might be connected via meaning postulates. For example, if there is a meaning postulate implying that a *sack* is a kind of *bag*, and that a *tabby* is a kind of *cat*, then sentences like (188) entail that a cat is out of a bag.

(188) #The tabby got out of the sack with a mouthful of salmon.

Pulman recognizes this problem and designs an indexing scheme to ensure that the right lexemes are present in the sentence. This weakens the claim that the approach is purely semantic/inferential, and it is still not powerful enough to exclude examples where both the right words and the right LF are present, but coincidentally:[7]

(189)*Yesterday our cat jumped out of the coal sack into the laundry bag.

This example shows that a more complicated indexing scheme would be required that links lexemes to the information they contribute to the LF.

It also seems that it is unnecessary and undesirable to have the full power of inference available. Since the examples in (190) match the antecedent of the relevant idiom rule and the right lexemes are present, there does not seem to be a way stating that these phrases do not have an idiomatic interpretation.

(190) a. #He boiled and then spilled the beans.

b. #He spilled the beans and some soda.

c. #The cat and the dog got out of the bag.

Any constraints on coordination there may be in the grammar are not applicable in this case, because the idiom rules are only applied after the rest of the grammar is already done with the utterance.

This approach also does not allow for syntactically ill-formed idioms, since these do not have a literal parse. Even sentences like (191) cannot be assigned a literal LF because there is no mass noun lexical entry for *shop*.

---

the antecedent is not literally true.

[7]The approach proposed in this dissertation does not have a problem with such examples because *bag* does not stand in the right relationship to *cat*, and *sack* does not have a *_bag_rel*.

(191) He closed up shop.

Furthermore, there is no way in this approach to represent the canonical forms of idioms, or to limit syntactic variation. Because there is no way to represent anything about the conventional syntactic properties of idioms, the possible range of variation for each idiom is predicted to follow entirely from other factors, e.g. pragmatics. This may work in some cases, but it cannot account for the full range of data. For example, it would not explain why active occurrences of *caught in the middle* do not have an idiomatic interpretation:

(192) a. When your son and daughter argue you will be caught in the middle.
b. #When your son and daughter argue they will catch you in the middle.

Finally, psycholinguistic evidence (e.g., McGlone et al. 1994) on the processing of idioms argues against an approach that requires literal meanings to be computed first. It is hard to see how idioms can be understood faster if they require additional processing during the inference stage. It is also hard to see what would explain the fact that the canonical forms of idioms are understood faster than other variants in an approach where there is no canonical form and no way of specifying anything about the form(s) of an idiom.

4.4 Summary

I summarize the preceding discussing in Table 4.1. + means the approach can deal with the problem, - means the approach cannot deal with the problem, -/+ means the approach can deal with the problem only when using additional methods, such as new features, functional uncertainty, or default unification, and +/- means that while the approach could handle the problem in its simple form, the introduction of the additional methods would make it unable to deal with the problem.[8] Of course, for some approaches some of these properties were hard to determine with certainty.

[8]Specifically, in the simple version of the approach subcategorizing for the syntax it is a problem to eliminate the literal meanings, but if this problem is avoided by referring to special idiomatic lexical entries, the problem arises of how to restrict their occurrence to idiomatic contexts.

Approaches	**Need for Phrasal Pattern**												**Variability**				
	literal meaning absent (2.2.1)	i_words not free (2.2.2)	no literal parse (2.2.3)	restricted flexibility (2.2.4)	canonical forms (2.2.5)	idiom families (2.2.6)	metaphorical mapping (2.2.7)	non-decomposable id. (2.2.8)	adjectives (2.2.9)	adjuncts (2.2.9)	no head (2.2.9)	psych. evidence (2.2.12)	inflection of comp. (2.3.1)	open slots (2.3.2)	modification (2.3.3)	passive (2.3.4)	across clauses (2.3.6)
Word-Level																	
multi-word lexeme (4.1.1)	+	+	+	+	+	–	–	+	+	+	+	+	–	–	–	–	–
subcat for phon. (4.1.2)	–	+	–	–	–	–	–	+	+	–	–	–	–	+	–	+	–
subcat for syntax (4.1.3)	–/+	+/–	–	–	–	–	–	+	–	–	–	–	+	+	–/+	+	–
LFG (4.1.4)	+	+	–	–	–	–	–	+	–	–	–	–	+	+	–	+	–
subcat for sem. (4.1.5)	+	–	+	–	–	–	–	–	–	–	–	–	+	+	+	+	–
GPSG (4.1.6)	+	–/+	–	–	–	+	–	–	–	–	–	–	+	+	+	+	–
Phrasal																	
fixed phonology (4.2.1)	–	+	+	+	+	–	–	+	+	+	+	+	–	–	–	–	–
fixed syntax (4.2.2)	+	–/+	+	+	+	+	–/+	+	+	+	+	+	+	–	–	–	–
TAG (4.2.3)	+	+	+	–	–	–	–	+	+	–	–	+	+	+	+	+	+
fixed semantics (4.2.4)	+	–/+	+	–/+	+	–/+	–/+	–/+	+	+	+	+	+	+	+	+	+
Jackendoff (4.2.5)	+	+	+	–	–/+	–	–	+	+	+	–/+	+	–	+	–	+	–
Other																	
Pulman (inference) (4.3)	+	+	–	–	–	+	+	+	+	+	+	–	+	+	+	+	+

Table 4.1: Idiom Approaches and their Problems

Not surprisingly, the word-level approaches tend to have problems with the type of data classified under the heading 'Need for Phrasal Pattern'. Some of these issues, such as representing canonical forms, idiom families, metaphorical mappings, and compatibility with psycholinguistic evidence, do not affect the generative capacity of the grammar, and such considerations may not be deemed relevant by everybody. But the other problems are not of this kind and cannot be ignored.

The phrasal approaches tend to do better with the items in this category, but tend to have problems with variability in idioms. It seems clear that a phrasal semantics approach is needed to account for the data. The particular version I propose, and an

account of how it deals with these data, is given in Chapter 5.

Chapter 5

The Constructional Approach

5.1 Introduction

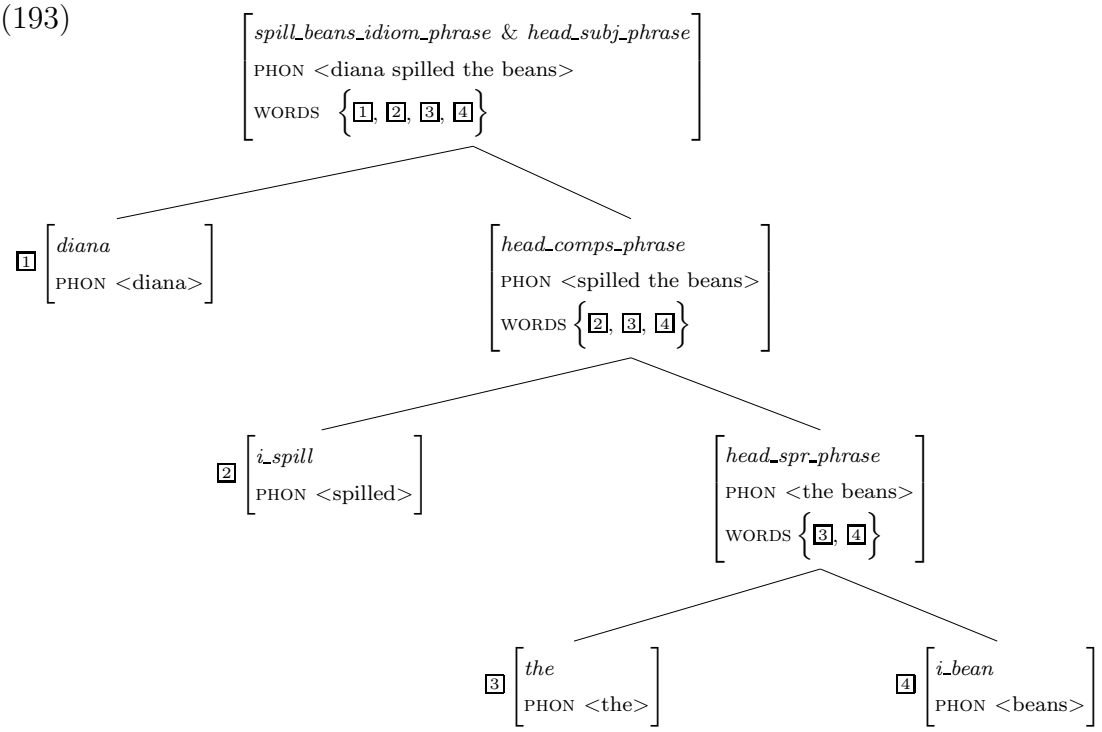
In this chapter I outline a new approach to idioms in the HPSG framework, building on earlier work by Copestake (1994). This constructional approach employs phrasal types that specify the semantic relationship between the idiomatic words involved. It does not require separate lexical entries for the words occurring in idioms, and treats the wide range of data discussed in Chapter 2 more successfully than previous alternatives. It can deal with semantically decomposable and non-decomposable idioms (Nunberg et al. 1994), and allows for the variability that some idioms exhibit, while being able to express what is fixed. It provides a solution to the problem of idioms occurring distributed over a main clause and a subordinate clause. The available psycholinguistic evidence is consistent with the approach, and it is intuitive to view an idiom like *spill the beans* as a whole phrasal unit, comprising the words *spill* and *beans* in their idiomatic meanings, without writing separate lexical entries for these idiomatic words.[1]

[1]As was discussed in Chapter 1, I use the following terminology in this dissertation. 'Idiomatic word' is used to mean 'word with a figurative meaning it has only as part of an idiom', e.g. *beans* meaning *secrets*. 'Lexical entry' is used to mean 'lexically specified representation of a word or phrase'.

5.2 The Constructional Approach

In the proposed constructional approach with partially fixed semantics, *phrases* have a set-valued[2] feature WORDS, which contains all the words in the phrase. More specifically, the items in this set are structure-shared with the terminal nodes of the syntactic tree,[3] as in (193). Note that because of this the COMPS list of transitive verbs in the WORDS set is not empty.

(193)



Apart from the added WORDS feature and the fact that two of the words in this sentence are idiomatic, this syntax tree is the same as the one for *Diana spilled the water* discussed in Chapter 1, although many details have been omitted for expository

[2]I call the feature set-valued in order to emphasize that the order of its elements is not significant, although this could perhaps be implemented by using a list in which the order is ignored. If it is not implemented as a list, one might think that it has to be a bag rather than a set, because there could be repeated elements. However, that is impossible in practice because the 'same' word does not play the same role in a sentence twice, so all the occurrences will be different from each other at least with respect to structure sharing.

[3]There are several possible ways to achieve this. One is to use the existing lexeme to word 'pumping' rules to introduce the WORDS feature, and then simply append the sets in all the syntactic rules. Another way is to have the syntactic rules append the WORDS sets of their phrasal daughters and add their non-phrasal daughters into the set.

reasons. One important difference is that idioms are represented as phrases with the WORDS feature partially instantiated, as in (194).

(194)
$$\left[\begin{array}{l} \textit{spill\_beans\_idiom\_phrase} \\ \text{WORDS} \left\{ \begin{array}{l} \left[\begin{array}{l} \textit{i\_word} \\ \ldots\text{LZT} \left\langle \left[\begin{array}{l} \textit{i\_spill\_rel} \\ \text{UND } \boxed{1} \end{array}\right] \right\rangle \end{array}\right], \\ \left[\begin{array}{l} \textit{i\_word} \\ \ldots\text{LZT} \left\langle \left[\begin{array}{l} \textit{i\_bean\_rel} \\ \text{INST } \boxed{1} \end{array}\right] \right\rangle \end{array}\right], \ \ldots \end{array} \right\} \end{array}\right]$$

In other words, parts of the idiom are specified as members of the set of WORDS of an idiomatic phrase. These idiomatic words can be related to the corresponding literal lexical entries using defaults:

(195)
$$\left[\begin{array}{l} \textit{spill\_beans\_idiom\_phrase} \\ \text{WORDS} \left\{ \begin{array}{l} \left[\begin{array}{l} \textit{i\_word} \\ \ldots\text{LZT} \left\langle \left[\begin{array}{l} \textit{i\_spill\_rel} \\ \text{UND } \boxed{1} \end{array}\right] \right\rangle \end{array}\right] \overset{<}{\sqcap} \left[\begin{array}{l} \textit{word} \\ \ldots\text{LZT } <\textit{\_spill\_rel}> \end{array}\right], \\ \left[\begin{array}{l} \textit{i\_word} \\ \ldots\text{LZT} \left\langle \left[\begin{array}{l} \textit{i\_bean\_rel} \\ \text{INST } \boxed{1} \end{array}\right] \right\rangle \end{array}\right] \overset{<}{\sqcap} \left[\begin{array}{l} \textit{word} \\ \ldots\text{LZT } <\textit{\_bean\_rel}> \end{array}\right], \ \ldots \end{array} \right\} \end{array}\right]$$

The signs included in the WORDS set are the ordinary literal lexical entries,[4] which have only their semantic relations overwritten in the description language by skeptical default unification of the kind proposed in Carpenter (1993) and extended to typed feature structures in Lascarides and Copestake (1999). More precisely, I am using their definition of asymmetric default unification on page 69. The description on the

[4]The type on the right side of the default unification symbol is *word*, not *spill*, in order to be compatible with passive forms (see Section 5.3.1).

left side of the $\overset{<}{\sqcap}$ symbol contains the strict information that gets augmented with all the non-conflicting information from the description on the right of the symbol. Since all this is done in the description language, i.e. in the description of lexical entries, the defaults do not persist, and there is no need for default unification in the underlying logic for combining signs. So the literal meanings are not present anywhere in the parse.

For the idiomatic meanings I use placeholders like *i_spill_rel* and *i_bean_rel*. This is because it is not relevant for the purpose of this dissertation to find out exactly what these idiomatic meanings are. In fact it may be impossible to do so, because for some speakers *i_spill_rel* may have the same meaning as *_reveal_rel*, for others it may mean the same thing as *_divulge_rel*, and for others it may not exactly correspond to the the meaning of any other verb of English. This notation is not intended to imply that there is only one idiomatic meaning that corresponds to each non-idiomatic word of English. In fact, when a word appears in more than one idiom, such as *strings* in *with no strings attached* and *pull strings*, it often does not have the same meaning, and in that case I would have to use different names, such as *i_string_1_rel* and *i_string_2_rel*. For background information about *rel*s, LZTs and other aspects of MRS, refer to Chapter 1.

For this type of idiom only the semantic relationships between the parts of the idiom are specified—the *beans* are the UNDERGOER (UND) of the *spill*ing (Davis 1996). This indirectly fixes the syntax to some extent, since the UNDERGOER is usually the head noun of the first NP on the COMPS list. But it is compatible with modification (*he spilled the politically charged beans*) and with idioms distributed over several clauses, since the parts stand in the specified semantic relationship. Syntactic variations like passivization also leave this semantic relationship unaffected. For a more detailed discussion of this, see Section 5.3.1.

The syntax is restricted further by the syntactic constraints on the individual words, for example their subcategorization requirements. These are usually the same as those on their non-idiomatic counterparts. If they are different they are explicitly mentioned in the definition of the idiom. So for example **spilled of the bean(s)* is blocked in the same way as **spilled of the juice* and **revealed of the secret(s)* are,

i.e. an *of*-complement is not compatible with the syntactic constraints in the lexical entries for the verbs *spill* and *reveal*, which are asking for an ordinary NP complement. These constraints are inherited by *i_spill*.

5.2.1 The Status of Idiomatic Words

The fact that idiomatic words cannot occur in their idiomatic meaning outside the idiom should be captured because these words do not have an existence independent of the idiomatic phrase they are a part of. By this I mean that there are no lexical entries for the individual idiomatic words, but only entries for the idiomatic phrases which contain them. This is conceptually clear and intuitive, and analogous to the explanation for why other items which are only parts of larger units, such as *synsem* objects, cannot occur by themselves. It is also implementable by modifying a parser to look through the set of WORDS of idiomatic phrases as well as through the set of ordinary lexical entries. When it is in the context of an idiom, the entire idiom phrase needs to be present for the parse to succeed. One way to implement this is suggested by Copestake (1993) in the context of phrasal transfer in shake-and-bake style machine translation (Whitelock 1992), where "idiom psorts[5] are only available to the parser/generator in the context of the relevant idiom". This does not quite capture the intuition that there should not be any separate representations for the idiomatic words, but it can be seen as one way to get the desired effect in an implementation.

Unfortunately, this intuition in its most natural form is not consistent with the current declarative logical frameworks, and probably could not be made consistent without additional machinery. Therefore I have worked out a formalization that works within the usual logical framework, although it does not capture the underlying intuition as straightforwardly.

There are two parts to the problem of why the intuitive approach is not consistent with the declarative logical framework. The first part is that it is necessary to define available words, i.e. words that objects of type *word* can resolve to, in such a way that we avoid the occurrence of *i_word*s with random combinations of phonology and

[5]In this approach lexical entries are psorts (pseudo-sorts) for various reasons. E.g., they do not need to introduce features.

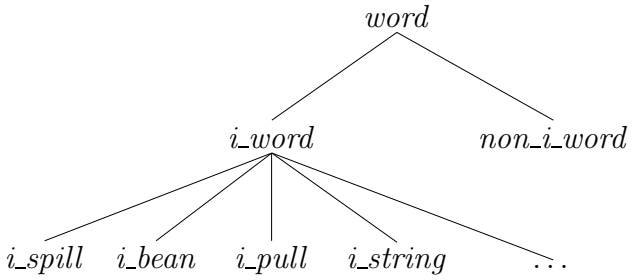
idiomatic meanings, e.g. the phonology *cat* and the meaning *i_spill_rel*.

One may wonder why this is necessary, while one does not have to worry about the occurrence of other impossible objects, such as for example *synsem* objects which have random combinations of meanings and HEAD values, resulting e.g. in an object with the meaning *growth_rel* which is also a *verb*. These kinds of random combinations are never a problem because *synsem* objects only occur as parts of *word*s, so only those *synsem* objects consistent with the lexical entries for those words can appear. However, at the word level it is necessary to say what all the possible subtypes (or instances) of the type *word* are,[6] so that the underspecified words contained in phrasal patterns do not get randomly resolved to contain irrelevant, incorrect information. This is true for both non-idiomatic and idiomatic words, although one may not realize it when one is used to treating lexical entries as subtypes as opposed to instances, because under a closed-world assumption this by definition results in a complete list of lexical entries being given to constrain what the possible *word*s are.

Ideally I would not want to have separate lexical entries for *i_word*s, to capture the intuition that they are not independent pairings of phonology and meaning. And unlike ordinary words I would like them to be instances and not types, to express that they have a different status and are not first-class citizens like ordinary words. But because it is necessary to have a constraint on *i_word* so that this type can only resolve to one of the *i_word*s implicitly declared in an idiom phrase, I have to make the *i_word*s types. This way I can refer to them when constraining the type *i_word* without having to give fully specified lexical entries for them at the word level. The link to the information about these *i_word*s, which is given as part of the definition of *i_phrase*s, is established via type inference. The type *word* and its subtypes can be seen in (196).

[6] I think of this as one way in which the formalism does not correspond to the reality of knowledge representation in humans. That is, I think that the concept of 'all the English idiomatic words I know' should be derivable implicitly—they are the ones I happen to have representations for—and it should be unnecessary to make an additional list of them. Specifically, if idiomatic words are thought of as instances, then it should be sufficient to say that these idiomatic words are instances of the type *i_word*, without also having to put an additional constraint on that type, mentioning what all its instances are. Maintaining this separate list is not only redundant, but would also get in the way if one tries to move away from representations as 'all-or-nothing', instead viewing them as having different levels of strength, corresponding to familiarity, frequency, etc.

(196)



Note that there are no constraints on these types given at this level. That is, the lexical entries for these *i_word*s are not written separately, but as part of the lexical entries for the *i_phrase*s. This presupposes that each individual *i_word* to be specified in this way occurs only in one idiom phrase, and it requires an addition to the description language to identify such specifications when they occur as part of larger constraints.

The second part of the problem is related to the fact that the *i_word*s exist as items that *word*s can resolve to, even though they are not listed separately as lexical entries on the word level. They now have to be prevented from turning up in non-idioms. This is not trivial because non-idiom phrases have to be consistent with containing *i_word*s, to allow sub-phrases like *some beans*, and larger phrases containing the idiom plus other material.

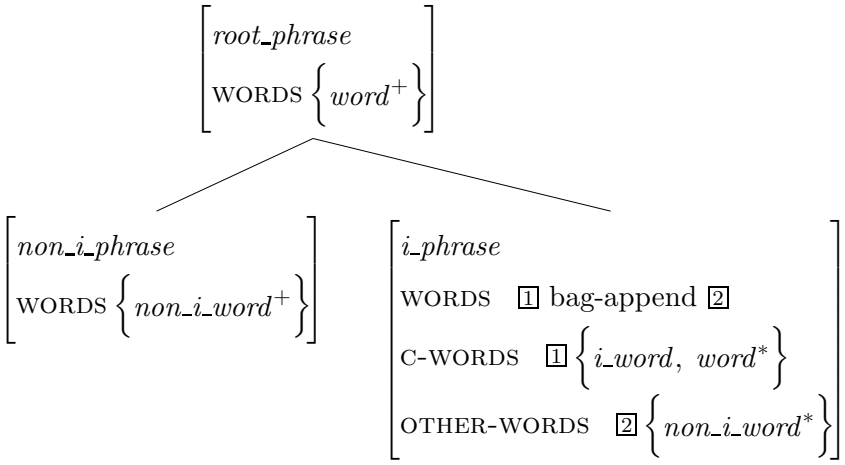
The solution to this problem is presented below. Intuitively, the idea is to keep a separate list of idiomatic words, and to add a constraint at the root level, i.e. the level of a complete utterance, to check that all the idioms are complete. The first step towards the formalization of this idea is to divide the hierarchy of *root_phrase*s, i.e. complete utterances, into *non_i_phrase*s and *i_phrase*s. The constraints on the types in this hierarchy state that *non_i_phrase*s only contain *non_i_word*s (words with a non-idiomatic meaning).[7]

*I_phrase*s have two additional features, C-WORDS (which is an abbreviation for CONSTRUCTIONAL-WORDS) and OTHER-WORDS, which together make up the set of

[7] Alternatively, there could be a similar constraint on the value of the LZT of *non_i_phrase*s, stating that this list contains only *non_i_rels*. But this would require complicating the semantics in a fashion analogous to the approach presented below, distinguishing between lists of idiomatic *rel*s and other *rel*s.

WORDS. This is necessary in order to be able to express the restriction that even in phrases containing idioms, only those *i_word*s are allowed which are part of that idiom. This is done by putting them in a separate complete set of C-WORDS, and constraining the set of the OTHER-WORDS in that phrase to have only members of type *non_i_word*. The value of the attribute WORDS is then defined as the result of appending the values of the attributes C-WORDS and OTHER-WORDS. The type hierarchy looks like (197).

(197)

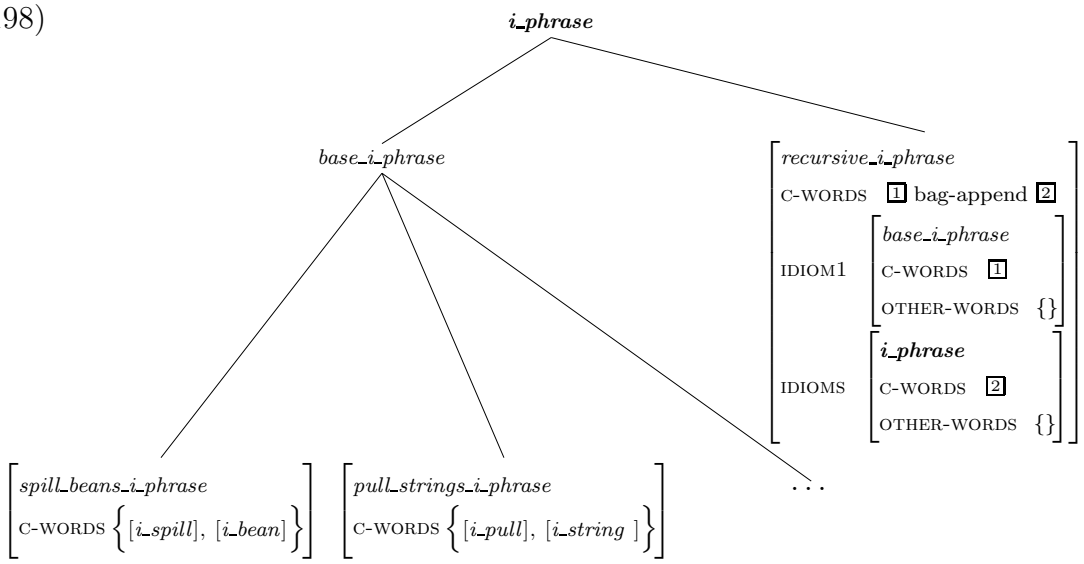


Note that the C-WORDS set can contain any number of *i_word*s because *i_word* is a subtype of *word*. The only way an *i_word* can be a member of the WORDS set, i.e. occur in a sentence, is by virtue of being in the C-WORDS set of an *i_phrase*. And the only way of getting into that set is by virtue of being introduced in the lexical entry of an *i_phrase*, which means that the other words that are part of that idiom are also present in this set of C-WORDS.

The *i_phrase* hierarchy is given in more detail in (198). It shows that *i_phrase*s can either be *base_i_phrase*s which contain only one idiom, or *recursive_i_phrase*s which contain more than one idiom.[8] In a *recursive_i_phrase* the values of the C-WORDS sets of the component idioms are appended. To avoid spurious ambiguity it is further stated that the *i_phrase*s in the recursion may not contain OTHER-WORDS, i.e. *non_i_word*s which are not part of the idiom.

[8]One might think that multiple inheritance would be enough to allow for more than one idiom, and this is true for the 'intuitive' approach in (194). However, if *i_word*s have an independent existence, multiple inheritance alone is not able to prevent these from occurring by themselves.

(198)



The boldface occurrences of *i_phrase* in the hierarchy in (198) highlight the recursion. That is, the value of the IDIOMS feature is of type *i_phrase*, which can be either a *base_i_phrase* or itself a *recursive_i_phrase*, so that any number of idioms can be handled via this constraint.

Note that this approach does not prevent idiomatic and non-idiomatic words from occurring together in a phrase. The only requirement is that if part of the idiom is present in a root phrase, the other parts have to be present as well, and stand in the appropriate semantic relationship.

The lexical entries for the idiomatic phrases now look like (199). In contrast to the simpler and more intuitive version in (195), the individual *i_word*s are types, and they are introduced in the C-WORDS attribute, to make the root-level constraints in (197) possible. Of course they are also in the set of WORDS, and the appending constraint in the root hierarchy in (197) makes sure that the set of C-WORDS and the set of OTHER-WORDS always add up to the complete set of WORDS in the utterance. Note that unlike in the version in (195), there are no 'dots' in the C-WORDS. This indicates that all the members of the set of C-WORDS are given.

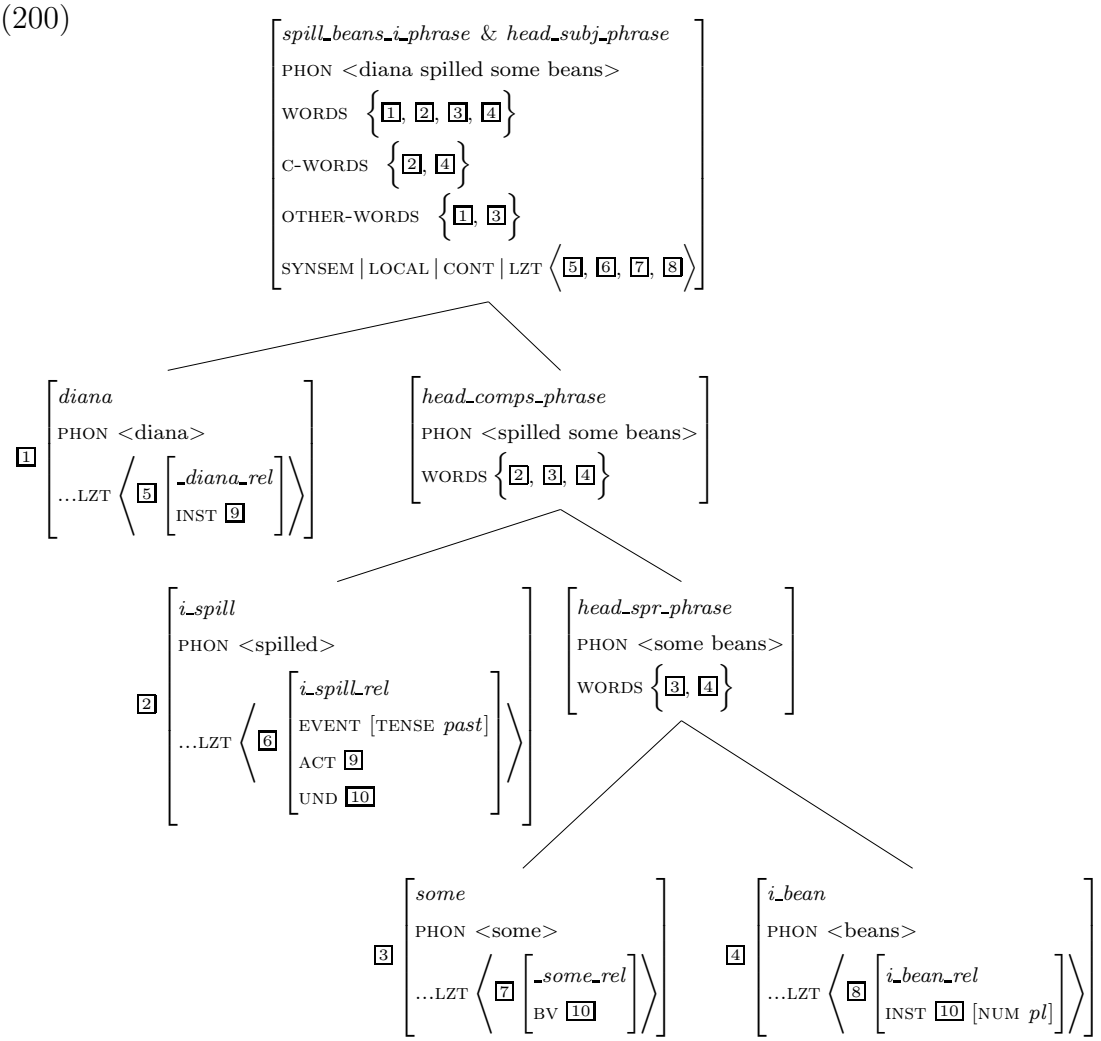
(199)
$$\left[\begin{array}{l} \textit{spill\_beans\_idiom\_phrase} \\ \textsc{c-words} \left\{ \begin{array}{l} \left[\begin{array}{l} \textit{i\_spill} \\ \ldots\textsc{lzt} \left\langle \left[\begin{array}{l} \textit{i\_spill\_rel} \\ \textsc{und} \ \boxed{1} \end{array}\right] \right\rangle \end{array}\right] \sqsubseteq \left[\begin{array}{l} \textit{word} \\ \ldots\textsc{lzt} \ <\textit{\_spill\_rel}> \end{array}\right], \\ \left[\begin{array}{l} \textit{i\_bean} \\ \ldots\textsc{lzt} \left\langle \left[\begin{array}{l} \textit{i\_bean\_rel} \\ \textsc{inst} \ \boxed{1} \end{array}\right] \right\rangle \end{array}\right] \sqsubseteq \left[\begin{array}{l} \textit{word} \\ \ldots\textsc{lzt} \ <\textit{\_bean\_rel}> \end{array}\right] \end{array} \right\} \end{array}\right]$$

To summarize, the idiomatic words are 'introduced' in the set of C-WORDS. At most phrasal levels the *i_word*s can freely mingle with non-idiomatic words. But at the root level a 'head count' is done to make sure that all and only those idiomatic words are present which originated in the set(s) of C-WORDS, making sure that the idiomatic words only appear when licensed by an *i_phrase*, so that no idiom is incomplete.

This formalization within the declarative logic captures the intuition that these *i_word*s are inextricably linked to the phrases they are a part of, and thereby linked to each other. It does not capture the intuition that the individual *i_word*s are not types and do not have any kind of existence outside the phrase, which is something that just cannot be expressed in this kind of logical framework. At least they have a somewhat different status from ordinary words—they are not listed explicitly in the lexicon as lexical items which are *i_word*s, but only as part of the lexical entries for the idioms which they are part of.

An example of an idiom not used in its canonical form, *Diana spilled some beans*, is given in (200):

(200)



As can be seen in this figure, *i_spill* and *i_bean* are elements of the C-WORDS set, so this utterance matches the constraints on the type *spill_beans_idiom_phrase* in (199). The OTHER-WORDS *diana* and *some* are also parts of the WORDS set, but are not C-WORDS. Apart from these added features, this syntax tree is the same as the one for *Diana spilled the water* discussed in Chapter 1.

It is not necessary that all words in the C-WORDS set be 'idiomatic words' in the sense of being of type *i_word*. When an idiom contains a word that has its literal meaning, e.g. *miss* in *miss the boat*, the lexical entry for this word can be listed as part of the C_WORDS set so that the mechanism described above will ensure that it is present when *i_boat* is present in an utterance:

(201)
$$\left[\begin{array}{l} \textit{miss\_boat\_idiom\_phrase} \\ \text{C-WORDS} \left\{ \begin{array}{l} \left[\begin{array}{l} \textit{word} \\ \ldots\text{LZT} \left\langle \left[\begin{array}{l} \textit{\_miss\_rel} \\ \text{UND } \boxed{1} \end{array}\right] \right\rangle \end{array}\right], \\ \left[\begin{array}{l} \textit{i\_boat} \\ \ldots\text{LZT} \left\langle \left[\begin{array}{l} \textit{i\_boat\_rel} \\ \text{INST } \boxed{1} \end{array}\right] \right\rangle \end{array}\right] \sqsubseteq \left[\begin{array}{l} \textit{word} \\ \ldots\text{LZT } <\textit{\_boat\_rel}> \end{array}\right] \end{array} \right\} \end{array}\right]$$

Similarly, when an idiomatic word occurs with the same figurative meaning in multiple idioms, it can be given a separate lexical entry. If this lexical entry is of type *i_word*, the mechanism described above ensures that it can only occur as part of the relevant idioms. One candidate for this type of treatment may be *ground* with the meaning in which it occurs in *lose ground*, *gain ground*, *make up ground*, and *recover (lost) ground*:

(202)
$$\left[\begin{array}{l} \textit{lose\_ground\_idiom\_phrase} \\ \text{C-WORDS} \left\{ \begin{array}{l} \left[\begin{array}{l} \textit{i\_lose} \\ \ldots\text{LZT} \left\langle \left[\begin{array}{l} \textit{i\_lose\_rel} \\ \text{UND } \boxed{1} \end{array}\right] \right\rangle \end{array}\right] \sqsubseteq \left[\begin{array}{l} \textit{word} \\ \ldots\text{LZT } <\textit{\_lose\_rel}> \end{array}\right], \\ \left[\begin{array}{l} \textit{i\_ground} \\ \ldots\text{LZT} \left\langle \left[\begin{array}{l} \textit{i\_ground\_rel} \\ \text{INST } \boxed{1} \end{array}\right] \right\rangle \end{array}\right] \end{array} \right\} \end{array}\right]$$

5.3 How the Approach Deals With the Problems

5.3.1 Variability

Variants Differing in Inflection

Variants differing in inflection are accounted for in this approach because the morphological information is part of the literal lexical entries referred to in the description of the idiomatic phrases, but it is left underspecified which of the inflected forms is part of the idiom. In most cases, the only property of a literal lexical entry referred to on the right side of the default unification symbol is its meaning, e.g., *_spill_rel*, which is consistent with any of the inflected forms of *spill*. When only particular forms can occur this can be stated explicitly, e.g., by saying that the word with the meaning *_bean_rel* also has to have the value *pl* for NUM.

Modification

Internal modification is possible for decomposable idioms involving verbs, since it does not interfere with the semantic relation between a verb and its arguments. If a part of an idiom is modified syntactically, the corresponding part of the idiomatic semantics is modified.

For example, the sentence *Diana spills the royal beans* is compatible with the constraints on *spill_beans_idiom_phrase*, because *royal* just adds an additional constraint but does not interfere with the relationship between *spill* and *beans*:

(203)
$$\left[\text{WORDS}\left\{\begin{bmatrix}\textit{i\_spill}\\ \text{...LZT}\left\langle\begin{bmatrix}\textit{i\_spill\_rel}\\ \text{UND } \boxed{1}\end{bmatrix}\right\rangle\end{bmatrix},\begin{bmatrix}\textit{the}\\ \text{...LZT}\left\langle\begin{bmatrix}\textit{the\_rel}\\ \text{BV } \boxed{1}\end{bmatrix}\right\rangle\end{bmatrix},\right.\right.$$
$$\left.\left.\begin{bmatrix}\textit{royal}\\ \text{...LZT}\left\langle\begin{bmatrix}\textit{\_royal\_rel}\\ \text{ARG } \boxed{1}\end{bmatrix}\right\rangle\end{bmatrix},\begin{bmatrix}\textit{i\_bean}\\ \text{...LZT}\left\langle\begin{bmatrix}\textit{i\_bean\_rel}\\ \text{INST } \boxed{1}\ [\text{NUM } \textit{pl}]\end{bmatrix}\right\rangle\end{bmatrix},\ \ldots\right\}\right]$$

Note that this applies only to intersective adjectives. If adjectives like *former* are given a standard scopal analysis, this predicts that *spill the former beans* does not

have an idiomatic interpretation.

Open Slots

Open slots like in *give NP some skin* are no problem because the missing bits, in this case the second argument of *give*, can just be left out. Unlike in a phonological approach it is not needed to 'make a connection' between the verb and its other complements.

(204)
$$\left[\begin{array}{l} \textit{give\_some\_skin\_idiom\_phrase} \\ \text{C-WORDS} \left\{\begin{array}{l} \left[\begin{array}{l} \textit{i\_give} \\ \ldots\text{LZT} \left\langle \left[\begin{array}{l} \textit{i\_give\_rel} \\ \text{UND } \boxed{1} \end{array}\right] \right\rangle \end{array}\right] \sqsubseteq \left[\textit{give\_ditrans}\right], \\ \left[\begin{array}{l} \textit{some} \\ \ldots\text{LZT} \left\langle \left[\begin{array}{l} \textit{some\_rel} \\ \text{BV } \boxed{1} \end{array}\right] \right\rangle \end{array}\right], \\ \left[\begin{array}{l} \textit{i\_skin} \\ \ldots\text{LZT} \left\langle \left[\begin{array}{l} \textit{i\_skin\_rel} \\ \text{INST } \boxed{1} \end{array}\right] \right\rangle \end{array}\right] \sqsubseteq \left[\begin{array}{l} \textit{word} \\ \ldots\text{LZT } <\textit{\_skin\_rel}> \end{array}\right] \end{array}\right\} \end{array}\right]$$

Note that the literal lexical entry for *give* used in this idiom is constrained to be ditransitive to ensure that it is not consistent with *give some skin to NP*, which does not have an idiomatic meaning.

Passive

Passive variants are accommodated, since the semantically underspecified representation is compatible among others with active and passive instantiations, and the phrasal idiom entries only refer to the semantic information of the literal words, and not the type of the lexical entries. The relevant parts of the lexical rule for passive are given in (205).

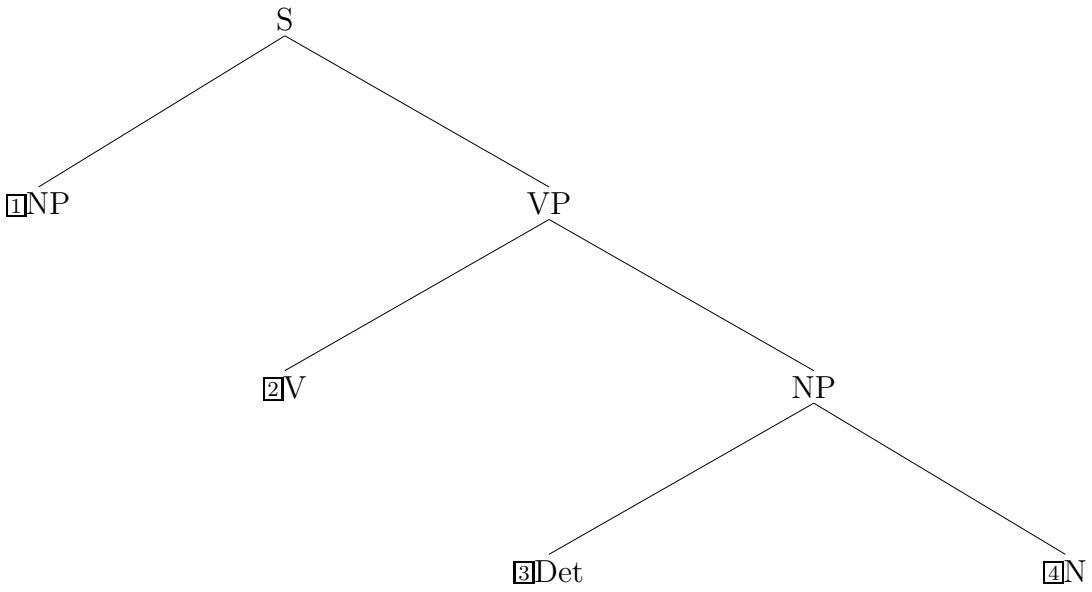
(205)
$$\left[\begin{array}{l} \textit{passive\_verb} \\ \textsc{phon}\ \textsc{psp}(\boxed{1}) \\ \ldots\textsc{vform}\ \textit{pass} \\ \ldots\textsc{subj} < [\ldots\textsc{cont}\ \boxed{2}] > \\ \ldots\textsc{comps} < \boxed{3} > \left(\oplus \left\langle \left[\ldots\textsc{lzt} \left\langle \left[\begin{array}{l} \textit{by\_rel} \\ \textsc{arg}\ \boxed{4} \end{array} \right] \right\rangle \right] \right\rangle \right) \\ \ldots\textsc{lzt}\ \boxed{5} \\ \textsc{stem} \left[\begin{array}{l} \textit{trans\_verb\_stem} \\ \textsc{phon}\ \boxed{1} \\ \ldots\textsc{subj} < [\ldots\textsc{index}\ \boxed{4}] > \\ \ldots\textsc{comps} < [\ldots\textsc{cont}\ \boxed{2}] > \oplus < \boxed{3} > \\ \ldots\textsc{lzt}\ \boxed{5} \end{array} \right] \end{array}\right]$$

Because the value of the LZT attribute is shared between the active and the passive verb, i.e. they have the same semantics, the ACTOR and UNDERGOER remain the same. In particular, the UNDERGOER is associated with the first complement of the active verb, which is the subject of the passive verb. (206) shows the passive form of the verb *spill*.

(206)
$$\left[\begin{array}{l}
\textit{passive_verb} \\
\textsc{phon}\ \textsc{psp}(\boxed{1}) \\
\ldots\textsc{subj}\ \left\langle \left[\ldots\textsc{cont}\ \boxed{2}\ [\textsc{index}\ \boxed{5}] \right] \right\rangle \\
\ldots\textsc{comps}\ <\ \boxed{3}\ >\ \left(\oplus \left\langle \left[\ldots\textsc{lzt}\ \left\langle \left[\begin{array}{l} \textit{by_rel} \\ \textsc{arg}\ \boxed{4} \end{array} \right] \right\rangle \right] \right\rangle \right) \\
\ldots\textsc{lzt}\ \boxed{6} \left\langle \left[\begin{array}{l} \textit{_spill_rel} \\ \textsc{act}\ \boxed{4} \\ \textsc{und}\ \boxed{5} \end{array} \right] \right\rangle \\
\textsc{stem} \left[\begin{array}{l} \textit{spill}\ \&\ \textit{trans_verb_stem} \\ \textsc{phon}\ \boxed{1}\ \langle \text{spill} \rangle \\ \ldots\textsc{subj}\ <\ [\textsc{index}\ \boxed{4}]\ > \\ \ldots\textsc{comps}\ <\ [\ldots\textsc{cont}\ \boxed{2}]\ >\ \oplus\ <\ \boxed{3}\ > \\ \ldots\textsc{lzt}\ \boxed{6} \end{array} \right]
\end{array}\right]$$

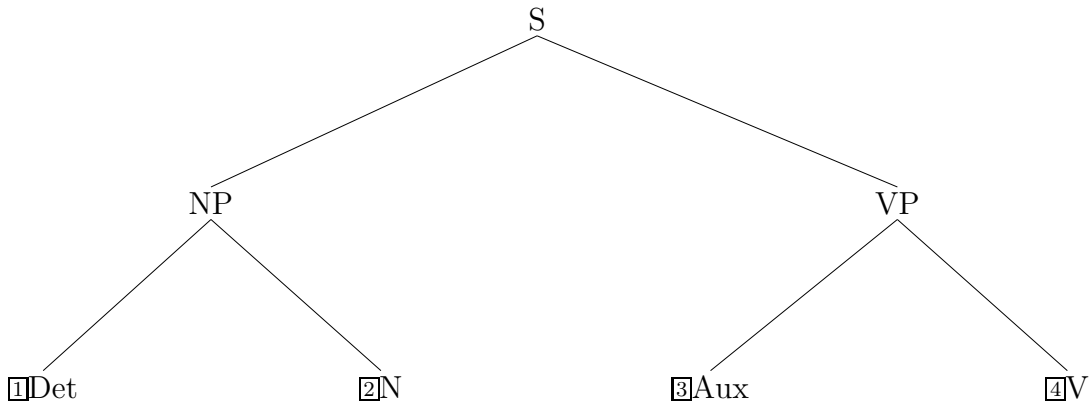
An example of an active sentence involving the idiom *spill the beans* can be seen in (207), and a corresponding passive example in (208).

(207)
$$\begin{bmatrix}
\textsc{phon} & \text{<diana spilled the beans>} \\
\textsc{words} & \left\{ \begin{bmatrix} \boxed{1} \begin{bmatrix} \textit{diana} \\ \textsc{phon} & \text{<diana>} \\ \ldots\textsc{lzt} & < \boxed{5} \ [\textsc{inst} \ \boxed{9}] > \end{bmatrix}, \boxed{2} \begin{bmatrix} \textit{i_spill} \\ \textsc{phon} & \text{<spilled>} \\ \ldots\textsc{lzt} & \left\langle \boxed{6} \begin{bmatrix} \textit{i_spill_rel} \\ \textsc{act} \ \boxed{9} \\ \textsc{und} \ \boxed{10} \end{bmatrix} \right\rangle \end{bmatrix}, \\ \boxed{3} \begin{bmatrix} \textit{the} \\ \textsc{phon} & \text{<the>} \\ \ldots\textsc{lzt} & < \boxed{7} \ [\textsc{bv} \ \boxed{10}] > \end{bmatrix}, \boxed{4} \begin{bmatrix} \textit{i_bean} \\ \textsc{phon} & \text{<beans>} \\ \ldots\textsc{lzt} & \left\langle \boxed{8} \begin{bmatrix} \textit{i_bean_rel} \\ \textsc{inst} \ \boxed{10} \end{bmatrix} \right\rangle \end{bmatrix} \end{bmatrix} \right\} \\
\textsc{synsem} \mid \textsc{local} \mid \textsc{cont} \mid \textsc{lzt} & \left\langle \boxed{5}, \boxed{6}, \boxed{7}, \boxed{8} \right\rangle
\end{bmatrix}$$



(208)

$$\begin{bmatrix}
\textsc{phon} & \text{<the beans were spilled>} \\
\textsc{words} & \left\langle \boxed{1}\begin{bmatrix} \textit{the} \\ \textsc{phon} & \text{<the>} \\ \ldots\textsc{lzt} & \left\langle \boxed{5}\begin{bmatrix} \textit{_the_rel} \\ \textsc{bv} & \boxed{9} \end{bmatrix} \right\rangle \end{bmatrix}, \boxed{2}\begin{bmatrix} \textit{i_bean} \\ \textsc{phon} & \text{<beans>} \\ \ldots\textsc{lzt} & \left\langle \boxed{6}\begin{bmatrix} \textit{i_bean_rel} \\ \textsc{inst} & \boxed{9} \end{bmatrix} \right\rangle \end{bmatrix}, \right. \\
& \left. \boxed{3}\begin{bmatrix} \textit{be} \\ \textsc{phon} & \text{<were>} \\ \ldots\textsc{subj} & \boxed{8} \end{bmatrix}, \boxed{4}\begin{bmatrix} \textit{i_spill} \\ \textsc{phon} & \text{<spilled>} \\ \ldots\textsc{subj} & \boxed{8} \\ \ldots\textsc{lzt} & \left\langle \boxed{7}\begin{bmatrix} \textit{i_spill_rel} \\ \textsc{und} & \boxed{9} \end{bmatrix} \right\rangle \end{bmatrix} \right\rangle \\
\textsc{synsem}\,|\,\textsc{local}\,|\,\textsc{cont}\,|\,\textsc{lzt} & \left\langle \boxed{5}, \boxed{6}, \boxed{7} \right\rangle
\end{bmatrix}$$



Raising Constructions

Some idioms can participate in raising constructions:

(209) The hatchet now appears to have been buried for good.

This is an example of subject raising, and *buried* is also passivized. Subject raising verbs like *appear* equate their complement's subject with their own subject:

(210)
$$\begin{bmatrix} \textit{subject\_raising\_verb} \\ \text{SYNSEM} \mid \text{LOCAL} \mid \text{CAT} \begin{bmatrix} \text{SUBJ } \boxed{1} \\ \text{COMPS} < [\ldots\text{SUBJ } \boxed{1}] > \end{bmatrix} \end{bmatrix}$$

Because the verb *buried* is passive, its subject's INDEX is the UNDERGOER. The CONTENT of its subject corresponds to the CONTENT of the complement of the active verb, because the whole subjects are equated. So the subject of the phrasal complement of *appear* corresponds to the subject of *appear*:

(211)
$$\begin{bmatrix} \textit{subject\_raising\_verb} \\ \text{PHON } <\text{appear}> \\ \text{SYNSEM} \mid \text{LOCAL} \begin{bmatrix} \text{CAT} \begin{bmatrix} \text{SUBJ } \boxed{1} \left\langle \begin{bmatrix} \ldots\text{INDEX } \boxed{2} \\ \ldots\text{LZT} \left\langle \begin{bmatrix} \textit{i\_hatchet\_rel} \\ \text{INST } \boxed{2} \end{bmatrix}, \ldots \right\rangle \end{bmatrix} \right\rangle \\ \text{COMPS} \left\langle \begin{bmatrix} \ldots\text{SUBJ } \boxed{1} \\ \ldots\text{INDEX } \boxed{3} \\ \ldots\text{LZT} \left\langle \begin{bmatrix} \textit{i\_bury\_rel} \\ \text{UND } \boxed{2} \end{bmatrix}, \ldots \right\rangle \end{bmatrix} \right\rangle \end{bmatrix} \\ \text{CONT} \mid \text{LZT} \left\langle \begin{bmatrix} \textit{appear\_rel} \\ \text{ARG } \boxed{3} \end{bmatrix} \right\rangle \end{bmatrix} \end{bmatrix}$$

The resulting sentence matches the constraint that the *hatchet* is the UNDERGOER of *bury* in the representation for *bury the hatchet* in (212).

(212)
$$\left[\begin{array}{l} \textit{bury\_hatchet\_idiom\_phrase} \\ \textsc{c-words} \left\{ \begin{array}{l} \left[\begin{array}{l} \textit{i\_bury} \\ \ldots\textsc{lzt} \left\langle \left[\begin{array}{l} \textit{i\_bury\_rel} \\ \textsc{und} \ \boxed{1} \end{array}\right] \right\rangle \end{array}\right] \sqcap^{\leq} \left[\begin{array}{l} \textit{word} \\ \ldots\textsc{lzt} <\textit{\_bury\_rel}> \end{array}\right], \\ \left[\begin{array}{l} \textit{i\_hatchet} \\ \ldots\textsc{lzt} \left\langle \left[\begin{array}{l} \textit{i\_hatchet\_rel} \\ \textsc{inst} \ \boxed{1} \end{array}\right] \right\rangle \end{array}\right] \sqcap^{\leq} \left[\begin{array}{l} \textit{word} \\ \ldots\textsc{lzt} <\textit{\_hatchet\_rel}> \end{array}\right] \end{array} \right\} \end{array}\right]$$

Control Constructions

With a few idioms in which the complement refers to an animate entity, control (equi) constructions are also possible:

(213) The piper wants to be paid.

Unlike raising verbs, control verbs do not share the whole CONTENT, but just the INDEX between their own subject and their complement's subject:

(214)
$$\left[\begin{array}{l} \textit{subject\_control\_verb} \\ \textsc{synsem}\,|\,\textsc{local}\,|\,\textsc{cat} \left[\begin{array}{l} \textsc{subj} < [\ldots\textsc{index} \ \boxed{1}] > \\ \textsc{comps} \left\langle \left[\ldots\textsc{subj} < [\ldots\textsc{index} \ \boxed{1}] > \right] \right\rangle \end{array}\right] \end{array}\right]$$

In this example, *paid* is again passivized, so its subject's INDEX is its UNDERGOER.

(215)
$$\left[\begin{array}{l} \textit{subject\_control\_verb} \\ \textsc{phon} \ <\text{want}> \\ \textsc{synsem}\,|\,\textsc{local} \left[\begin{array}{l} \textsc{cat} \left[\begin{array}{l} \textsc{subj} \left\langle \left[\ldots\textsc{cont} \left[\begin{array}{l} \textsc{index} \ \boxed{1} \\ \textsc{lzt} \left\langle \left[\begin{array}{l} \textit{i\_piper\_rel} \\ \textsc{inst} \ \boxed{1} \end{array}\right] \ldots \right\rangle \end{array}\right] \right] \right\rangle \\ \textsc{comps} \left\langle \left[\begin{array}{l} \ldots\textsc{subj} < \left[\ldots\textsc{index} \ \boxed{1}\right] > \\ \ldots\textsc{cont} \left[\begin{array}{l} \textsc{index} \ \boxed{2} \\ \textsc{lzt} \left\langle \left[\begin{array}{l} \textit{i\_pay\_rel} \\ \textsc{und} \ \boxed{1} \end{array}\right] \ldots \right\rangle \end{array}\right] \end{array}\right] \right\rangle \end{array}\right] \\ \textsc{cont}\,|\,\textsc{lzt} \left\langle \left[\begin{array}{l} \textit{want\_rel} \\ \textsc{act} \ \boxed{1} \\ \textsc{arg} \ \boxed{2} \end{array}\right] \right\rangle \end{array}\right] \end{array}\right]$$

The fact that only the INDEX is shared is not a problem for the analysis proposed here, because the right semantic relationship specified in the representation for *pay the piper* in (216) is still present.

(216)
$$\left[\begin{array}{l} \textit{pay\_piper\_idiom\_phrase} \\ \textsc{c-words} \left\{\begin{array}{l} \left[\begin{array}{l} \textit{i\_pay} \\ \ldots\textsc{lzt} \left\langle \left[\begin{array}{l} \textit{i\_pay\_rel} \\ \textsc{und} \ \boxed{1} \end{array}\right] \right\rangle \end{array}\right] \sqsubseteq \left[\begin{array}{l} \textit{word} \\ \ldots\textsc{lzt} \ <\textit{\_pay\_rel}> \end{array}\right], \\ \left[\begin{array}{l} \textit{i\_piper} \\ \ldots\textsc{lzt} \left\langle \left[\begin{array}{l} \textit{i\_piper\_rel} \\ \textsc{inst} \ \boxed{1} \end{array}\right] \right\rangle \end{array}\right] \sqsubseteq \left[\begin{array}{l} \textit{word} \\ \ldots\textsc{lzt} \ <\textit{\_piper\_rel}> \end{array}\right] \end{array}\right\} \end{array}\right]$$

Topicalization

The lexical entries for the idiom phrases are also consistent with topicalized utterances like (217).

(217) The other beans, she will probably spill later.

This is because the *head_filler_phrase* specifies that the entire LOCAL information including all of the semantics is shared between the filler and the gap. The relevant constraints on *head_filler_phrase* can be seen in (218).

(218)
$$\begin{bmatrix} \textit{head\_filler\_phrase} \\ \text{HEAD-DTR} \left[\ldots\text{SLASH} < \boxed{1} > \right] \\ \text{NON-HEAD-DTR} \left[\text{SYNSEM} \middle| \text{LOCAL} \ \boxed{1} \right] \end{bmatrix}$$

The gap is something that originated on the COMPS list[9] and therefore for transitive verbs like *spill* its INDEX is the UNDERGOER, as can be seen in (219).

(219)
$$\begin{bmatrix} \textit{head\_filler\_phrase} \\ \text{HEAD-DTR} \begin{bmatrix} \ldots\text{COMPS} < \ > \\ \ldots\text{SLASH} \left\langle \boxed{1} \left[\ldots\text{INDEX} \ \boxed{2} \right] \right\rangle \\ \ldots\text{LZT} \left\langle \begin{bmatrix} \textit{\_spill\_rel} \\ \text{UND} \ \boxed{2} \end{bmatrix}, \ \ldots \right\rangle \end{bmatrix} \\ \text{NON-HEAD-DTR} \left[\text{SYNSEM} \middle| \text{LOCAL} \ \boxed{1} \right] \end{bmatrix}$$

Distribution Over Several Clauses

The McCawley data, i.e. idioms distributed over a main clause and a relative clause, can be handled since the parts again stand in the specified semantic relationship. For example, in the feature structure for the sentence in (220), the LZT value is compatible with that of the idiom representation in (221).

(220) The strings that Kim pulled got Chris the job.

[9] Whether it ended up in SLASH via an argument realization principle or via a lexical rule is not relevant here.

(221)
$$\left[\begin{array}{l} \textit{pull\_strings\_idiom\_phrase} \\ \textsc{c-words} \left\{ \begin{array}{l} \left[\begin{array}{l} \textit{i\_pull} \\ \ldots\textsc{lzt} \left\langle \left[\begin{array}{l} \textit{i\_pull\_rel} \\ \textsc{und} \ \boxed{1} \end{array}\right] \right\rangle \end{array}\right] \sqsubseteq \left[\begin{array}{l} \textit{word} \\ \ldots\textsc{lzt} \ <\textit{\_pull\_rel}> \end{array}\right], \\ \left[\begin{array}{l} \textit{i\_string} \\ \ldots\textsc{lzt} \left\langle \left[\begin{array}{l} \textit{i\_string\_rel} \\ \textsc{inst} \ \boxed{1} \end{array}\right] \right\rangle \end{array}\right] \sqsubseteq \left[\begin{array}{l} \textit{word} \\ \ldots\textsc{num} \quad \textit{pl} \\ \ldots\textsc{lzt} \ <\textit{\_string\_rel}> \end{array}\right] \end{array} \right\} \end{array}\right]$$

To understand why this is the case, it is necessary to show briefly how relative clauses are treated. Intuitively, the semantic variable of the missing complement in the relative clause is equated with that of the noun modified by the relative clause. The relevant constraints on the type *filler_head_relative_clause* can be seen in (222).

(222)
$$\left[\begin{array}{l} \textit{filler\_head\_relative\_clause} \\ \ldots\textsc{mod} \ [\ldots\textsc{index} \ \boxed{2}] \\ \textsc{head-dtr} \left[\ldots\textsc{slash} \left\{\boxed{3}\right\}\right] \\ \textsc{non-head-dtr} \left[\begin{array}{l} \ldots\textsc{local} \ \boxed{3} \ [\ldots\textsc{index} \ \boxed{2}] \\ \ldots\textsc{rel} \left\{\boxed{2}\right\} \end{array}\right] \end{array}\right]$$

These constraints ensure that the value of SLASH is structure-shared with the LOCAL value of the relative pronoun, and that its semantic INDEX is structure-shared with the INDEX of the MOD value of the relative clause.

The constraints on the type *head_adjunct_relative_phrase* can be seen in (223).

(223)
$$\left[\begin{array}{l} \textit{head\_adjunct\_relative\_phrase} \\ \ldots\textsc{index} \ \boxed{1} \\ \textsc{head-dtr} \ [\ldots\textsc{index} \ \boxed{1}] \\ \textsc{non-head-dtr} \left[\begin{array}{l} \textit{filler\_head\_relative\_clause} \\ \ldots\textsc{mod} \ [\ldots\textsc{index} \ \boxed{1}] \end{array}\right] \end{array}\right]$$

These constraints make sure that the value of the INDEX of the NON-HEAD-DTR's MOD feature are the same as that of the HEAD-DTR.

Because the missing complement in the phrase *kim pulled* originated on the COMPS list and was put into SLASH, the INDEX of the element in SLASH, which is unified with the relative pronoun, is the UNDERGOER of *pull*. As can be seen in (222), the relative pronoun's INDEX is shared with that of the noun being modified by the relative clause. The lexical entries for nouns equate the values of their INDEX feature and their INST. The result of the interaction of all these constraints can be seen in (224).

(224)
$$\left[\begin{array}{l} \textit{head\_adjunct\_relative\_phrase} \\ \text{PHON } \langle\text{strings that kim pulled}\rangle \\ \text{HEAD-DTR } \left[\begin{array}{l} \text{PHON } \langle\text{strings}\rangle \\ \ldots\text{INDEX } \boxed{1} \\ \ldots\text{LZT } \left\langle \left[\begin{array}{l} \textit{i\_string\_rel} \\ \text{INST } \boxed{1} \end{array}\right] \right\rangle \end{array}\right] \\ \text{NON-HEAD-DTR } \left[\begin{array}{l} \textit{filler\_head\_relative\_clause} \\ \text{PHON } \langle\text{that kim pulled}\rangle \\ \ldots\text{MOD } [\ldots\text{INDEX } \boxed{1}] \\ \text{HEAD-DTR } \left[\begin{array}{l} \text{PHON } \langle\text{kim pulled}\rangle \\ \ldots\text{LZT } \left\langle \left[\begin{array}{l} \textit{\_kim\_rel} \\ \text{INST } \boxed{3} \end{array}\right], \left[\begin{array}{l} \textit{i\_pull\_rel} \\ \text{ACT } \boxed{3} \\ \text{UND } \boxed{1} \end{array}\right] \right\rangle \\ \ldots\text{SLASH } \left\{ \boxed{2} [\ldots\text{INDEX } \boxed{1}] \right\} \end{array}\right] \\ \text{NON-HEAD-DTR } \left[\begin{array}{l} \text{PHON } \langle\text{that}\rangle \\ \ldots\text{LOCAL } \boxed{2} [\ldots\text{INDEX } \boxed{1}] \\ \ldots\text{REL } \left\{ \boxed{1} \right\} \end{array}\right] \end{array}\right] \end{array}\right]$$

So the phrase *the strings that kim pulled* can occur as the subject of *got Chris the job* even though *the strings* by itself cannot, because *strings* is licensed to occur in its idiomatic meaning by the presence of *pull* in the relative clause.

In the approach developed in Sag (1997) even in 'bare' or '*that*-less' relative clauses only the INDEX is shared between the missing element in the relative clause and the noun being modified by the relative clause:

(225)
$$\begin{bmatrix} \textit{non\_wh\_relative\_clause} \\ \ldots\textsc{mod}\ [\ldots\textsc{index}\ \boxed{1}] \\ \textsc{head-dtr}\ \left[\ldots\textsc{slash}\ \left\{\ldots\textsc{index}\ \boxed{1}\right\}\right] \end{bmatrix}$$

Pronominal Reference

The approach predicts that it is not possible to refer to the noun from a non-decomposable idiom with a pronoun, because there is no meaning associated with such nouns. This seems to be a correct prediction:

(226) *He kicked the bucket, and he kicked it yesterday.

For decomposable idioms the approach can handle certain types of pronominal reference and ellipsis. Because idiomatic words like *beans* are associated with a meaning in these idioms, they can be referred to with pronouns in the usual way, whether in the same sentence or across sentence boundaries. An example of this type is given in (227).

(227) No soap opera worth its bubbles would spill all the beans in one episode if it could dribble them out over many.

There were no corpus examples of VP ellipsis as in (228), but if they occur they can be handled for the same reasons.

(228) I thought the beans would be spilled, but they weren't.

In another type of pronominal reference, an 'idiomatic word' seems to appear by itself as in (229):

(229) Eventually she spilled all the beans. But it took her a few days to spill them all.

In this type, the pronominal reference is not the problem, but the fact that idiomatic *spill* is not 'licensed' may be a problem. However, many speakers use metaphorical verbs like *spill* with non-idiomatic nouns, as in *spill secrets*. These speakers presumably have an independent lexical entry for *spill* with its metaphorical meaning, so that examples like (229) should also be acceptable for these speakers.

It remains to be seen what exactly the restrictions are on this phenomenon,[10] and how it might be formalized if examples like (229) occur and are accepted by speakers who reject *spill secrets*.

5.3.2 Properties Shared between Literal and Idiomatic Words

The approach predicts that words in idioms have the same morphology and syntax as their literal counterparts by default, and can occur in all inflected variants unless specified otherwise. It also has the virtue of coming closer to explaining how some further semantic properties seem to be inherited from literal meanings (Nunberg 1977). For example, *kick the bucket* can only be used to describe instantaneous instances of dying:

(230) *He kicked the bucket slowly, over a period of several hours.

This would be predicted if it is assumed that events are typed and classified according to this kind of information, and that in this case the only thing that is default-overwritten is the *rel*. So parts of the non-idiomatic information can 'survive', as long as they are compatible with the idiomatic meaning. But in most cases these kinds of properties are those of the idiomatic event, as can be seen in (231). Literal *spilling* tends to be involuntary and tends to happen fast, while these properties are not necessarily present in the idiomatic meaning.

(231) He spilled the beans slowly, over the course of several days.

This can be described by overwriting additional parts of the default information from the literal lexical entry. It may not be necessary to list these constraints explicitly, because the idiomatic meaning can determine the syntax and semantics in the same way that literal verb meanings usually do. For example, *i_spill_rel* may inherit such constraints from a more general type, that *_divulge_rel* also inherits from.

[10] There are two reasons why it would be hard to do a corpus study of this. The first reason is that the queries would produce a much larger percentage of irrelevant matches to sort through. The second problem is that many tools for corpus research have built-in options for searching 'within one sentence', but not 'within two adjacent sentences'. But specifying an arbitrary limit like 'within 25 words' would miss relevant examples while producing an even larger number of irrelevant matches.

5.3.3 No Literal Interpretation

Items like *close up shop* conform to existing syntactic patterns of English, but do not have a literal parse because of the properties of the words involved, in this case because *shop* is not a mass noun. Other examples are *tend shop*, and *set foot*. These can be handled by default-overwriting the property in question in the description of the lexical entries for the idiomatic phrases:

(232)
$$\left[\begin{array}{l} \textit{tend\_shop\_idiom\_phrase} \\ \textsc{c-words} \left\{ \begin{array}{l} \left[\begin{array}{l} \textit{i\_tend} \\ \ldots\textsc{lzt} \left\langle \left[\begin{array}{l} \textit{i\_tend\_rel} \\ \textsc{und} \ \boxed{1} \end{array}\right] \right\rangle \end{array}\right] \sqsubseteq \left[\begin{array}{l} \textit{word} \\ \ldots\textsc{lzt} <\textit{\_tend\_rel}> \end{array}\right], \\ \left[\begin{array}{l} \textit{i\_shop} \\ \ldots\textsc{count} \ \textit{-} \\ \ldots\textsc{lzt} \left\langle \left[\begin{array}{l} \textit{i\_shop\_rel} \\ \textsc{inst} \ \boxed{1} \end{array}\right] \right\rangle \end{array}\right] \sqsubseteq \left[\begin{array}{l} \textit{word} \\ \ldots\textsc{lzt} <\textit{\_shop\_rel}> \end{array}\right] \end{array} \right\} \end{array}\right]$$

5.3.4 Restricting the Flexibility of Idioms

It is also possible to restrict the flexibility of idioms, for example for idioms that never occur actively like *be caught in the middle*. This type of constraint can be stated by specifying the VFORM of the verb as in (233). Only passive verbs will satisfy this constraint (see Section 5.3.1).

(233)
$$\left[\begin{array}{l} \textit{caught\_in\_the\_middle\_idiom\_phrase} \\ \textsc{c-words} \left\{ \begin{array}{l} \left[\begin{array}{l} \textit{i\_catch} \\ \ldots\textsc{lzt} \left\langle \boxed{1} \left[\begin{array}{l} \textit{i\_catch\_rel} \\ \textsc{event} \ \boxed{2} \end{array}\right] \right\rangle \end{array}\right] \sqsubseteq \left[\begin{array}{l} \textit{word} \\ \ldots\textsc{lzt} <\textit{\_catch\_rel}> \\ \ldots\textsc{vform} \ \textit{pass} \end{array}\right], \ \ldots \end{array} \right\} \end{array}\right]$$

Other types of restrictions on the flexibility, such as occurrence only in active but not in passive sentences, can also be stated in terms of the VALENCE or DTRS attributes. This is discussed in Section 5.3.9 in the context of canonical forms.

5.3.5 Idiom Families

It is possible to exploit the non-idiomatic hierarchies of words or semantic relations for the purpose of describing idiom families, since these are accessible from within the description of the idiom. The literal lexical entries including their *rel*s are available in the description, and *rel*s are types which can be organized in a type hierarchy. So, for *throw someone to the wolves/lions* a supertype *large_carnivorous_animal_rel* could be used, as can be seen from the description of that idiom in (234).[11]

(234)
$$\left[\begin{array}{l} \textit{throw\_to\_dangerous\_animals\_idiom\_phrase} \\ \textsc{c-words} \left\{ \begin{array}{l} \left[\begin{array}{l} \textit{i\_throw} \\ \ldots\textsc{lzt} \left\langle \left[\begin{array}{l} \textit{i\_throw\_rel} \\ \textsc{event}\ \boxed{2} \\ \textsc{arg2}\ \boxed{1} \end{array}\right] \right\rangle \end{array}\right] \sqsubseteq \left[\begin{array}{l} \textit{word} \\ \ldots\textsc{lzt} <\textit{\_throw\_rel}> \end{array}\right], \\ \left[\begin{array}{l} \textit{to} \\ \ldots\textsc{lzt} \left\langle \left[\begin{array}{l} \textit{\_to\_rel} \\ \textsc{arg}\ \boxed{2} \\ \textsc{arg3}\ \boxed{1} \end{array}\right] \right\rangle \end{array}\right], \left[\begin{array}{l} \textit{the} \\ \ldots\textsc{lzt} \left\langle \left[\begin{array}{l} \textit{\_the\_rel} \\ \textsc{bv}\ \boxed{1} \end{array}\right] \right\rangle \end{array}\right], \\ \left[\begin{array}{l} \textit{i\_word} \\ \ldots\textsc{lzt} < [\textsc{inst}\ \boxed{1}] > \end{array}\right] \sqsubseteq \left[\begin{array}{l} \textit{noun} \\ \ldots\textsc{num}\ \textit{pl} \\ \ldots\textsc{lzt} <\textit{large\_carnivorous\_animal\_rel}> \end{array}\right] \end{array} \right\} \end{array}\right]$$

As argued in Chapter 2, this is a generalization that might be worth expressing because it probably influences which further variants are likely to be coined. But it is not strong enough by itself to predict the observed data, and it is necessary to list the conventionalized members of such families.

5.3.6 Locus for the Metaphorical Mapping

Another advantage of this approach is that there is one place that contains the metaphorical mapping—both the literal meanings the idiomatic meanings of all the words

[11] The LinGO grammar does not use such supertypes unless they are needed for grammatical reasons. Instead a disjunction could be used.

that are part of such a conventionalized metaphor are present in the same representation. There could be hierarchies of such mappings, with common metaphors making idioms which instantiate them easier to learn and remember. For example, such hierarchies could express the relationship between the idioms *break the mold* and *fit the mold*, and also help to explain how non-conventionalized uses of these metaphors like *shatter the mold*, *crack the mold*, and *break out of the mold* can be interpreted. This would require an addition to the formalism, as a system for hierarchies of description language constraints using this type of default has not been developed.

5.3.7 Locus for Semantics of Non-Decomposable Idioms

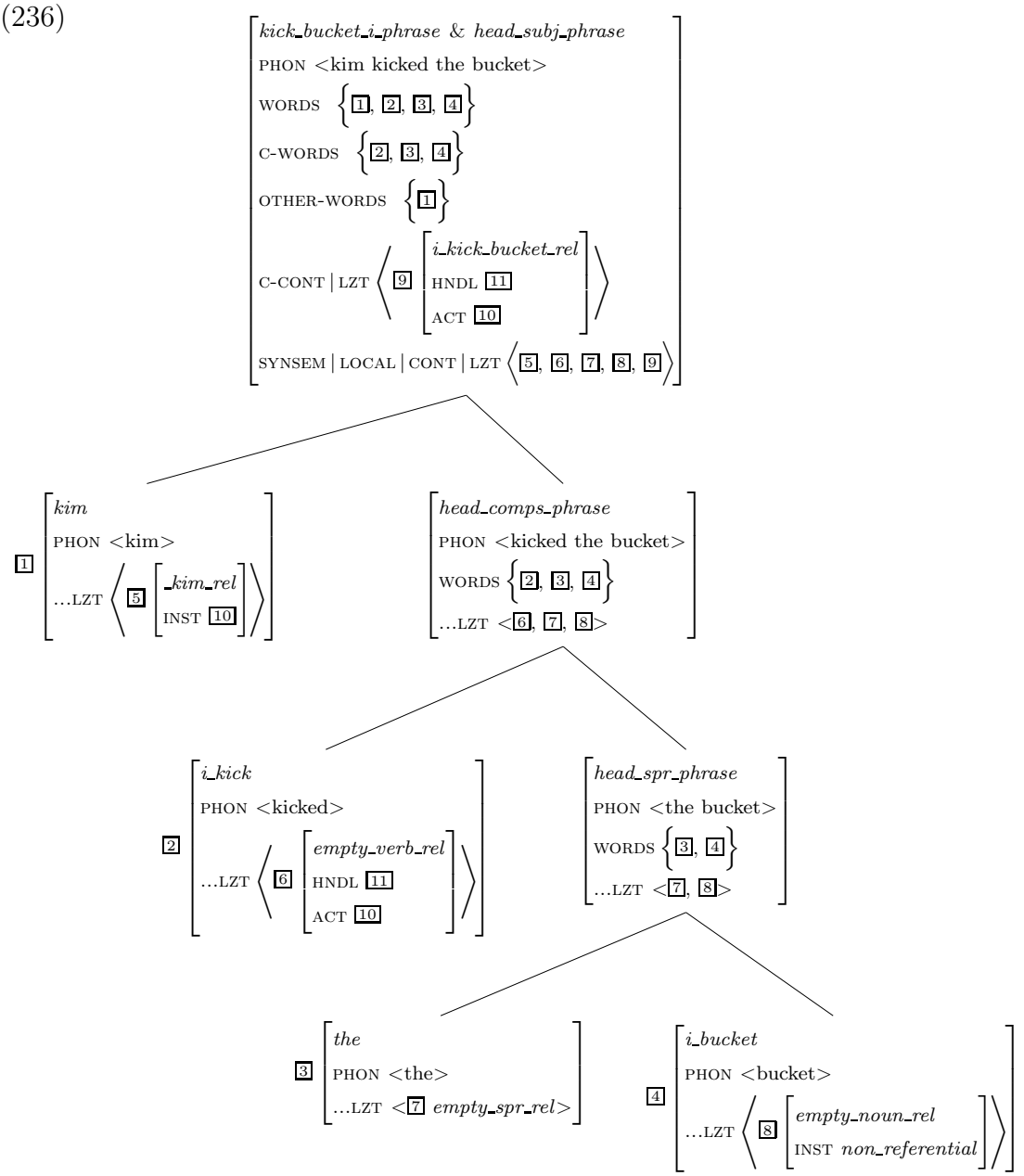
For non-decomposable idioms, the words *kick*, *the*, and *bucket* do not contribute to the meaning of the idiom, and these words are instead associated with an *empty_rel*,[12] and the whole idiomatic construction contributes an *i_kick_bucket_rel*. The feature C-CONT is used in the LinGO grammar to encode the semantic contribution of constructions, and the meaning of non-decomposable idioms can be seen as a special case of this. This has the advantage that the Semantics Principle applies as usual—the LZT of a phrase contains the *rels* from all the daughters plus those of the C-CONT. So there is no *i_bucket_rel*, as is appropriate for these idioms, which do not distribute their meaning over their syntactic parts.

[12]In order for the syntax and linking to work as usual, this *empty_rel* cannot be as 'empty' as one may think. For example they have to have HNDL attributes and there have to be subsorts of *empty_rel* for verbs, nouns, etc., so that for example an *empty_noun_rel* has an attribute INST as required by the grammar.

(235)
$$\left[\begin{array}{l} \textit{kick\_bucket\_idiom\_phrase} \\ \textsc{c-words} \left\{\begin{array}{l} \left[\begin{array}{l} \textit{i\_kick} \\ \ldots\textsc{subj} \left\langle \left[\ldots\textsc{index}\ \boxed{1}\right] \right\rangle \\ \ldots\textsc{comps} \left\langle \left[\ldots\textsc{index}\ \boxed{2}\right] \right\rangle \\ \ldots\textsc{lzt} \left\langle \left[\begin{array}{l} \textit{empty\_verb\_rel} \\ \textsc{hndl}\ \boxed{4} \\ \textsc{act}\ \boxed{1} \end{array}\right] \right\rangle \end{array}\right] \sqsubseteq \left[\begin{array}{l} \textit{word} \\ \ldots\textsc{lzt} <\textit{\_kick\_rel}> \end{array}\right] \\ \left[\begin{array}{l} \textit{i\_the} \\ \ldots\textsc{lzt} <\boxed{3}\ \textit{empty\_spr\_rel}> \end{array}\right] \sqsubseteq \left[\begin{array}{l} \textit{word} \\ \ldots\textsc{lzt} <\textit{\_the\_rel}> \end{array}\right], \\ \left[\begin{array}{l} \textit{i\_bucket} \\ \ldots\textsc{spr} \left\langle \left[\ldots\textsc{lzt} < \boxed{3} >\right] \right\rangle \\ \ldots\textsc{lzt} \left\langle \left[\begin{array}{l} \textit{empty\_noun\_rel} \\ \textsc{inst}\ \boxed{2}\ \textit{non\_referential} \end{array}\right] \right\rangle \end{array}\right] \sqsubseteq \left[\begin{array}{l} \textit{word} \\ \ldots\textsc{lzt} <\textit{\_bucket\_rel}> \end{array}\right] \end{array}\right\} \\ \textsc{c-cont} \mid \textsc{lzt} \left\langle \left[\begin{array}{l} \textit{i\_kick\_bucket\_rel} \\ \textsc{hndl}\ \boxed{4} \\ \textsc{act}\ \boxed{1} \end{array}\right] \right\rangle \end{array}\right]$$

As can be seen in (236), the syntactic structure of the sentence *Kim kicked the bucket* is the same as that of *Diana spilled the water*, and all the normal principles of syntactic and semantic combination apply in the same way. The main difference is that some of the items that end up on the LZT are *empty_rel*s, so they do not contribute anything to the meaning. Instead the meaning of the idiom is specified as part of the constructional meaning in C-CONT.

(236)



This approach predicts that it is impossible for *proverbial* or any other adjective to modify *bucket* semantically, because there is no *i_bucket_rel* and the *empty_noun_rel* does not have a referential index. But there is no syntactic problem with a *proverbial bucket*, because the exact location of *bucket* in the NP is not specified. This is because it is stated only that the INDEX of *bucket* is the INDEX of the NP complement of *kick*. The mechanism for achieving the unusual wide scope of *proverbial* in *he*

kicked the proverbial bucket, i.e. *proverbially (speaking), he died*, is not explored here. But such a mechanism is needed independently for examples like *an occasional sailor walked in*, which means *occasionally a sailor walked in*. Further constraints may be needed to exclude other adjectives that should be able to modify the whole sentence semantically. However, corpus examples like *who made them kick their respective buckets*, which means *who made them die, respectively*, suggest that these should not be excluded in principle.

The fact that non-decomposable idioms do not passivize and are not involved in topicalization is captured, because *the bucket* is specified to be the complement of *kick* syntactically. That is, the phrase is constrained to contain the verb *kick* actually taking *the bucket* as its complement. The lexical entry for the passive form *kicked* is not consistent with this constraint, because its only complement is an optional by-phrase (see Section 5.3.1). And the constraint in (235) is also not consistent with *the bucket* being the topicalized element in a *head_filler_phrase*, because that element is not found on the COMPS list, but is an element of SLASH. Note that the elements of WORDS are structure-shared with the leaf nodes of the syntactic tree, where the COMPS requirements have not been discharged. The same kind of specification would not have the same effect in a word level approach, since lexical rules can apply there.

5.3.8 More than Head-Argument Relationships

Adjectives and Specifiers

Fixed adjectives and specifiers pose no problem because they can just be put in the right semantic relationship to what they specify or modify. A fixed adjective can be seen in (237).

(237)
$$\left[\begin{array}{l} \textit{bark\_up\_wrong\_tree\_idiom\_phrase} \\ \text{C-WORDS} \left\{\begin{array}{l} \left[\begin{array}{l} \textit{wrong} \\ \ldots\text{LZT} \left\langle \left[\begin{array}{l} \textit{wrong\_rel} \\ \text{ARG } \boxed{1} \end{array}\right] \right\rangle \end{array}\right], \\ \left[\begin{array}{l} \textit{i\_tree} \\ \ldots\text{LZT} \left\langle \left[\begin{array}{l} \textit{i\_tree\_rel} \\ \text{INST } \boxed{1} \end{array}\right] \right\rangle \end{array}\right] \sqsubseteq \left[\begin{array}{l} \textit{word} \\ \ldots\text{LZT } <\textit{\_tree\_rel}> \end{array}\right] \end{array}\right\} \end{array}\right]$$

Depending on one's treatment of the copula, this representation might also be consistent with *bark up the tree which is wrong*, which does not have an idiomatic reading. In that case it is possible to add a further constraint stating that the adjective is attributive, i.e. [PRED -]. Note that I am assuming here that *wrong* is not an idiomatic word, as it seems to have its usual meaning. It occurs in all the definitions given in my two idioms dictionaries (*follow the wrong course (of action)*, *make the wrong choice*, and *ask the wrong person*). Note that the analysis would be analogous for adjectives with an idiomatic meaning.

A fixed specifier can be seen in (238).

(238)
$$\left[\begin{array}{l} \textit{give\_some\_skin\_idiom\_phrase} \\ \text{C-WORDS} \left\{\begin{array}{l} \left[\begin{array}{l} \textit{some} \\ \ldots\text{LZT} \left\langle \left[\begin{array}{l} \textit{some\_rel} \\ \text{BV } \boxed{1} \end{array}\right] \right\rangle \end{array}\right], \\ \left[\begin{array}{l} \textit{i\_skin} \\ \ldots\text{LZT} \left\langle \left[\begin{array}{l} \textit{i\_skin\_rel} \\ \text{INST } \boxed{1} \end{array}\right] \right\rangle \end{array}\right] \sqsubseteq \left[\begin{array}{l} \textit{word} \\ \ldots\text{LZT } <\textit{\_skin\_rel}> \end{array}\right] \end{array}\right\} \end{array}\right]$$

Adverbs and Adjuncts

Adjuncts can be handled because they can be specified to semantically modify the event. This can be seen for an adverb in (239) and for a PP adjunct in (240).

(239)
$$\left[\begin{array}{l} \textit{put\_mildly\_idiom\_phrase} \\ \text{C-WORDS} \left\{\begin{array}{l} \left[\begin{array}{l} \textit{i\_put} \\ \ldots\text{LZT} \left\langle \left[\begin{array}{l} \textit{i\_put\_rel} \\ \text{EVENT } \boxed{1} \end{array}\right] \right\rangle \end{array}\right] \sqsubseteq \left[\begin{array}{l} \textit{word} \\ \ldots\text{LZT } <\textit{\_put\_rel}> \end{array}\right], \\ \left[\begin{array}{l} \textit{i\_mildly} \\ \ldots\text{LZT} \left\langle \left[\begin{array}{l} \textit{i\_mildly\_rel} \\ \text{ARG } \boxed{1} \end{array}\right] \right\rangle \end{array}\right] \sqsubseteq \left[\begin{array}{l} \textit{word} \\ \ldots\text{LZT } <\textit{\_mildly\_rel}> \end{array}\right] \end{array}\right\} \end{array}\right]$$

(240)
$$\left[\begin{array}{l} \textit{skate\_on\_thin\_ice\_idiom\_phrase} \\ \text{C-WORDS} \left\{\begin{array}{l} \left[\begin{array}{l} \textit{i\_skate} \\ \ldots\text{LZT} \left\langle \left[\begin{array}{l} \textit{i\_skate\_rel} \\ \text{HNDL } \boxed{2} \\ \text{EVENT } \boxed{3} \end{array}\right] \right\rangle \end{array}\right] \sqsubseteq \left[\begin{array}{l} \textit{word} \\ \ldots\text{LZT } <\textit{\_skate\_rel}> \end{array}\right], \\ \left[\begin{array}{l} \textit{i\_on} \\ \ldots\text{LZT} \left\langle \left[\begin{array}{l} \textit{i\_on\_rel} \\ \text{HNDL } \boxed{2} \\ \text{ARG } \boxed{3} \\ \text{ARG3 } \boxed{1} \end{array}\right] \right\rangle \end{array}\right] \sqsubseteq \left[\begin{array}{l} \textit{word} \\ \ldots\text{LZT } <\textit{\_on\_rel}> \end{array}\right], \\ \left[\begin{array}{l} \textit{i\_thin} \\ \ldots\text{LZT} \left\langle \left[\begin{array}{l} \textit{i\_thin\_rel} \\ \text{HNDL } \boxed{4} \\ \text{ARG } \boxed{1} \end{array}\right] \right\rangle \end{array}\right] \sqsubseteq \left[\begin{array}{l} \textit{word} \\ \ldots\text{LZT } <\textit{\_thin\_rel}> \end{array}\right], \\ \left[\begin{array}{l} \textit{i\_ice} \\ \ldots\text{LZT} \left\langle \left[\begin{array}{l} \textit{i\_ice\_rel} \\ \text{HNDL } \boxed{4} \\ \text{INST } \boxed{1} \end{array}\right] \right\rangle \end{array}\right] \sqsubseteq \left[\begin{array}{l} \textit{word} \\ \ldots\text{LZT } <\textit{\_ice\_rel}> \end{array}\right] \end{array}\right\} \end{array}\right]$$

No Verb or Other Head

It is not a problem that some idioms do not involve a fixed verb. In idioms like *get/set/start/keep/have the ball rolling* the relevant relationship between the parts, e.g., the location of an underspecified event, can be expressed without stating to which verbal *rel* they belong.[13]

(241)
$$\left[\begin{array}{l} \textit{ball\_rolling\_idiom\_phrase} \\ \text{C-WORDS} \left\{ \begin{array}{l} \left[\ldots\text{LZT} \left\langle \left[\begin{array}{l}\textit{relation}\\ \text{ARG3 } \boxed{1}\\ \text{ARG4 } \boxed{2}\end{array}\right\rangle\right] \\ \left[\begin{array}{l}\textit{i\_ball}\\ \ldots\text{LZT} \left\langle \left[\begin{array}{l}\textit{i\_ball\_rel}\\ \text{INST } \boxed{1}\end{array}\right\rangle\right]\end{array}\right] \sqsubseteq \left[\begin{array}{l}\textit{word}\\ \ldots\text{NUM } \textit{sg}\\ \ldots\text{LZT } <\textit{\_ball\_rel}>\end{array}\right] \\ \left[\begin{array}{l}\textit{i\_roll}\\ \ldots\text{LZT} \left\langle \left[\begin{array}{l}\textit{i\_roll\_rel}\\ \text{EVENT } \boxed{2}\end{array}\right\rangle\right]\end{array}\right] \sqsubseteq \left[\begin{array}{l}\textit{word}\\ \ldots\text{VFORM } \textit{prp}\\ \ldots\text{LZT } <\textit{\_roll\_rel}>\end{array}\right] \end{array} \right\} \end{array}\right]$$

If a preposition is involved, as in idioms like *up the creek without a paddle*, *butterflies in one's stomach*, or *cat among the pigeons*, the preposition can establish a link between its two arguments as in (242).

[13] Something more needs to be said about the type of semantic relation to rule out #*see/remember/consider the ball rolling*.

(242)
$$\left[\begin{array}{l} \textit{cat\_among\_pigeons\_idiom\_phrase} \\ \text{C-WORDS} \left\{\begin{array}{l} \left[\begin{array}{l}\textit{i\_cat}\\ \ldots\text{LZT} \left\langle \left[\begin{array}{l}\textit{i\_cat\_rel}\\ \text{INST } \boxed{1}\end{array}\right]\right\rangle\end{array}\right] \sqsubseteq \left[\begin{array}{l}\textit{word}\\ \ldots\text{LZT } <\textit{\_cat\_rel}>\end{array}\right], \\ \left[\begin{array}{l}\textit{among}\\ \ldots\text{LZT} \left\langle \left[\begin{array}{l}\textit{among\_rel}\\ \text{ARG1 } \boxed{1}\\ \text{ARG2 } \boxed{2}\end{array}\right]\right\rangle\end{array}\right], \\ \left[\begin{array}{l}\textit{the}\\ \ldots\text{LZT} \left\langle \left[\begin{array}{l}\textit{the\_rel}\\ \text{BV } \boxed{2}\end{array}\right]\right\rangle\end{array}\right], \\ \left[\begin{array}{l}\textit{i\_pigeon}\\ \ldots\text{LZT} \left\langle \left[\begin{array}{l}\textit{i\_pigeon\_rel}\\ \text{INST } \boxed{2}\end{array}\right]\right\rangle\end{array}\right] \sqsubseteq \left[\begin{array}{l}\textit{word}\\ \ldots\text{NUM } \textit{pl}\\ \ldots\text{LZT } <\textit{\_pigeon\_rel}>\end{array}\right] \end{array}\right\} \end{array}\right]$$

5.3.9 Canonical Forms

It possible to express which variant, e.g., with a particular choice of specifier or syntactic structure, is the canonical form of an idiom, by making it a subtype of the more general representation of the idiom, while allowing for further non-lexicalized variations. For example, if one wants to build a psycholinguistically plausible model capturing the data described in Chapter 3, it is necessary to state that the canonical forms are the forms in which these idioms are most likely to occur. This might also be useful in an implementation for speeding up parsing, and for preventing the generation of output which is not 'idiomatic' in the non-technical sense.[14]

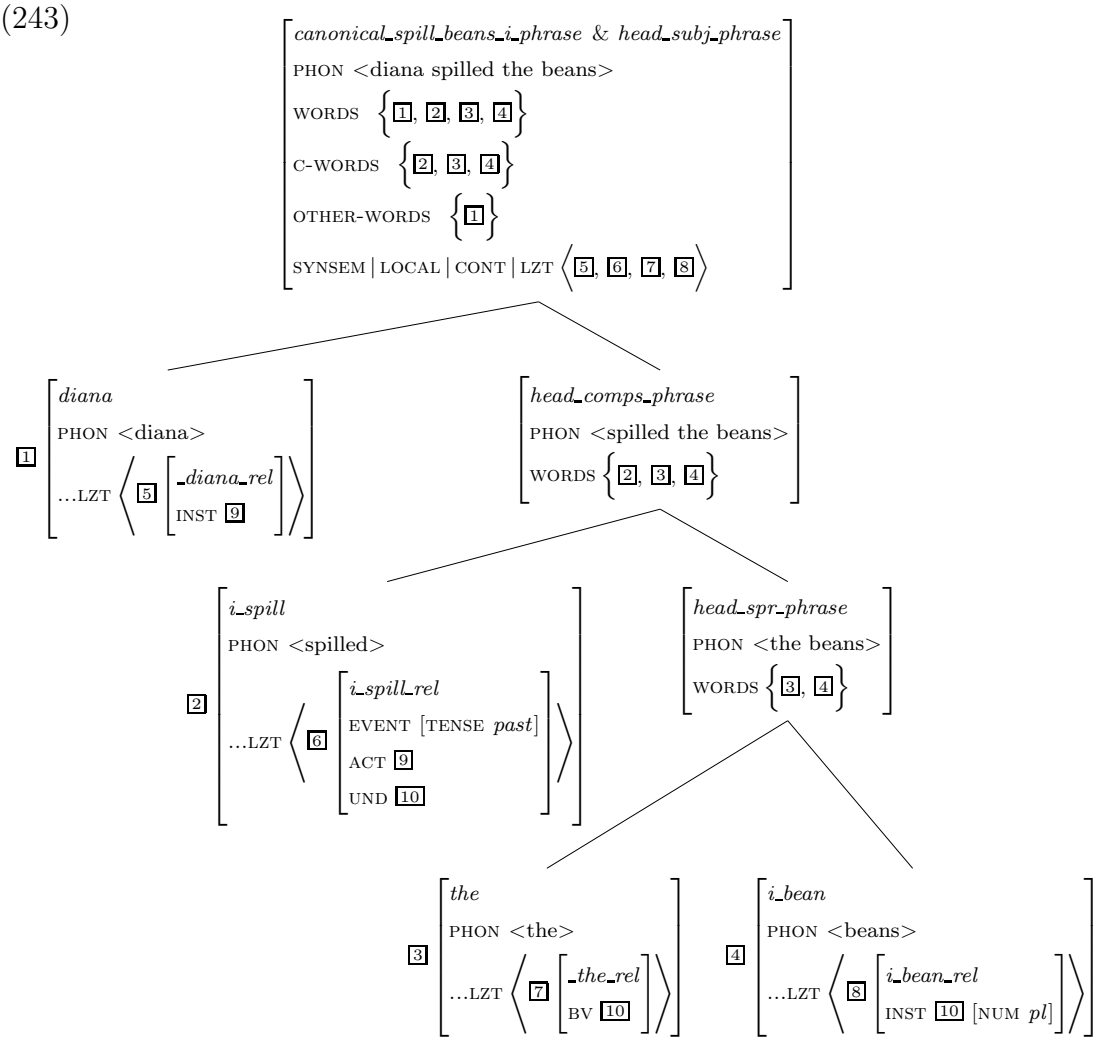
Since the canonical form is the one that is heard and used most frequently and is

[14]In Chapter 3 it was shown that 25% of the data can only be parsed if the grammar allows for variability. But from the point of view of generation it is probably safest to generate idioms only in their canonical form, in order to avoid violating any constraints on variability which were not foreseen by the grammar writer, or which are hard to formalize because they are of a pragmatic nature and require world knowledge.

probably learned first,[15] it would not surprising if it has a representation of its own in the minds of speakers, even though they have realized that the meaning of that particular idiom is decomposable and that it occurs in a variety of constructions. For further discussion of this see Section 9.1.

An example of an idiom used in its canonical form, *Diana spilled the beans*, is given in (243):

(243)



Note that in this example, unlike in (200), the specifier is *the*, and it is one of the

[15]Of course there is a 25% chance that a learner might hear a non-canonical form of a particular idiom first. But learning a new word or idiom does not necessarily happen instantaneously when it is first encountered. Nevertheless, there may still be some speakers who temporarily learn a non-canonical variant before they encounter the canonical one.

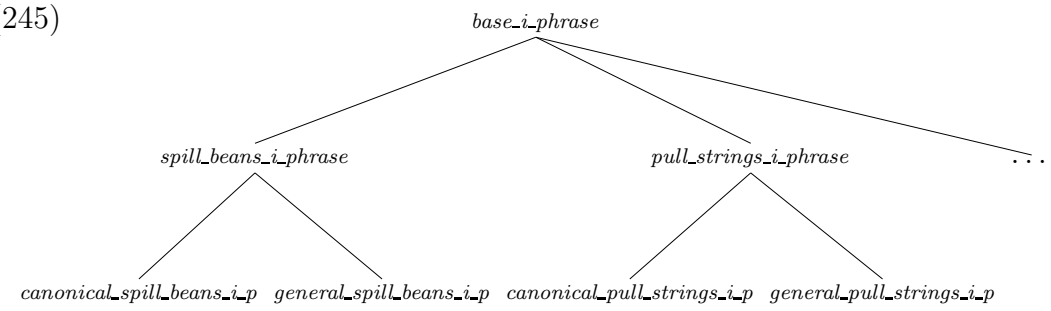
C-WORDS. Otherwise the two examples are not very different from each other because in the non-canonical use in (200) the *beans* happened to be the complement of *spill*.

An example of a lexical entry for the canonical form of *spill the beans* can be found in (244). All the constraints on the canonical form are shown, including the ones inherited from the more general entry for *spill_beans_idiom_phrase*.

(244)
$$\left[\begin{array}{l} \textit{canonical\_spill\_beans\_i\_p} \\ \\ \text{C-WORDS} \left\{\begin{array}{l} \left[\begin{array}{l} \textit{i\_spill} \\ \text{...COMPS} < [\text{...INDEX } \boxed{1}] > \\ \text{...LZT} \left\langle \left[\begin{array}{l} \textit{i\_spill\_rel} \\ \text{UND } \boxed{1} \end{array}\right] \right\rangle \end{array}\right] \leq\sqcap \left[\begin{array}{l} \textit{word} \\ \text{...LZT} < \textit{\_spill\_rel} > \end{array}\right], \\ \left[\begin{array}{l} \textit{the} \\ \text{SYNSEM } \boxed{2} \left[\begin{array}{l} \text{...SPEC } [\text{...INDEX } \boxed{1}] \\ \text{...LZT} < \textit{\_the\_rel} > \end{array}\right] \end{array}\right], \\ \left[\begin{array}{l} \textit{i\_bean} \\ \text{...SPR} < \boxed{2} > \\ \text{...LZT} \left\langle \left[\begin{array}{l} \textit{i\_bean\_rel} \\ \text{INST } \boxed{1} \end{array}\right] \right\rangle \end{array}\right] \leq\sqcap \left[\begin{array}{l} \textit{word} \\ \text{...NUM } \textit{pl} \\ \text{...LZT} < \textit{\_spill\_rel} > \end{array}\right] \end{array}\right\} \end{array}\right]$$

Note that it is necessary to create another subtype *general_spill_beans_i_p* to allow the more general type to be used productively.[16] There are no further constraints stated on this type, because they are identical to the ones on its immediate supertype.

(245)

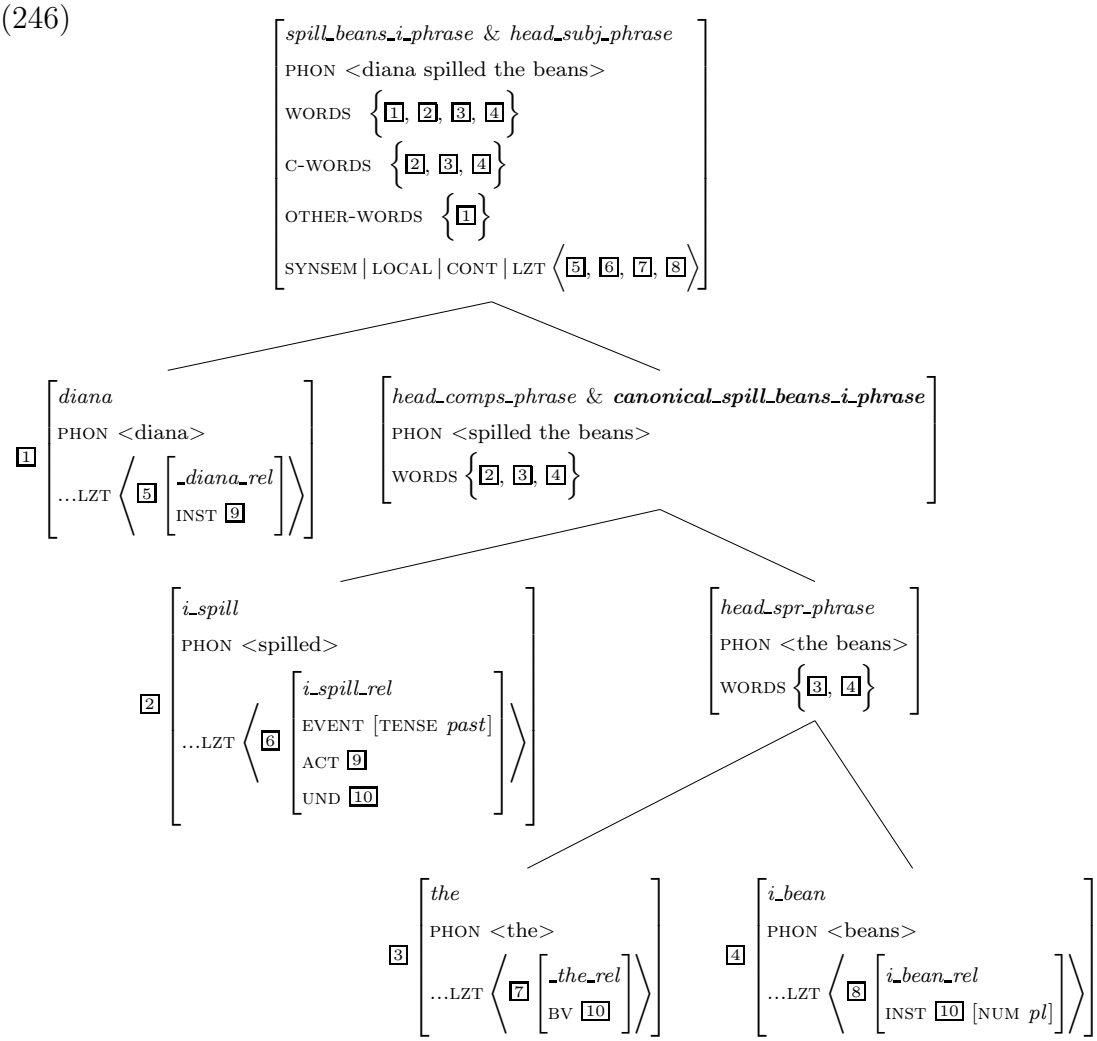


[16] I find this somewhat counterintuitive and instead prefer to think of the generalizations themselves being marked as productive.

This is how a canonical form analysis would probably be implemented by a grammar writer. That is, only the minimal amount of information would be given that is required to make sure that the phrase matches only the canonical form *spill the beans* modulo inflection of *spill*. For example, it is not necessary to view this canonical phrase as a VP—as long as its parts are constrained syntactically or semantically to make up the correct VP, there is no problem if the constraint is consistent with larger phrases containing this VP. In such an implementation the representation for the canonical form would be used either to attach frequency information, or to allow only the canonical form to be used in generation, to avoid unwanted creative, non-idiomatic sounding sentences being generated.

However, if one wants to be psycholinguistically plausible, it is actually quite possible that people store the complete representation of the phrase with most of its feature values and structure sharings, so that the grammar does not need to do any work assembling this VP. I am not going to show this complete phrase in a figure because it would contain too much information—it would correspond to the complete VP subtree of (246). For the same reason a grammar writer would probably want to avoid writing a lexical entry like that. But it would be possible to store these kinds of phrases automatically, using a mechanism similar to that proposed in Neumann (1997), but deciding which parsed phrases to store as lexical entries for the canonical forms of idioms. Neumann also developed a mechanism for using such phrases for parsing and generation. The phrasal templates are added as a passive edge to the chart parser/generator's agenda, and passive edges with a larger span are given higher priority. Note that in this case this *canonical_spill_beans_i_phrase* would be a *head_comps_phrase*, i.e., not compatible with being at the root level. However, this does not cause a problem for the analysis given above, because at the root level the whole utterance can still be of type *spill_beans_i_phrase*, licensing the occurrence of the constructional words, as shown in (246).

(246)



Note that (246) is the same as (243) except for the location of the type *canonical_spill_beans_i_phrase*, highlighted in bold face.

One problem with this approach to canonical forms is that canonical occurrences of idioms can be parsed in two ways—using the representation for the canonical form, or using the more general representation that is consistent with varied as well as canonical occurrences. This is analogous to the situation found with established compounds which have not semantically drifted and could therefore also be derived productively. For example, the compound *vending machine* is fully compositional in that it means *machine for vending*, and compounds with a 'purpose' relation can be formed productively in English (*graphing software*). Yet *vending machine* is also

clearly an established word of English, while productively coined words like *selling machine* or *vending device* are not. This is one of the ways in which morphological and idiomatic 'productivity' is not of the same nature as full syntactic productivity. That is, new words and varied idioms are often noticed as unusual and have a 'new' feel to them, which is not the case for ordinary syntax.

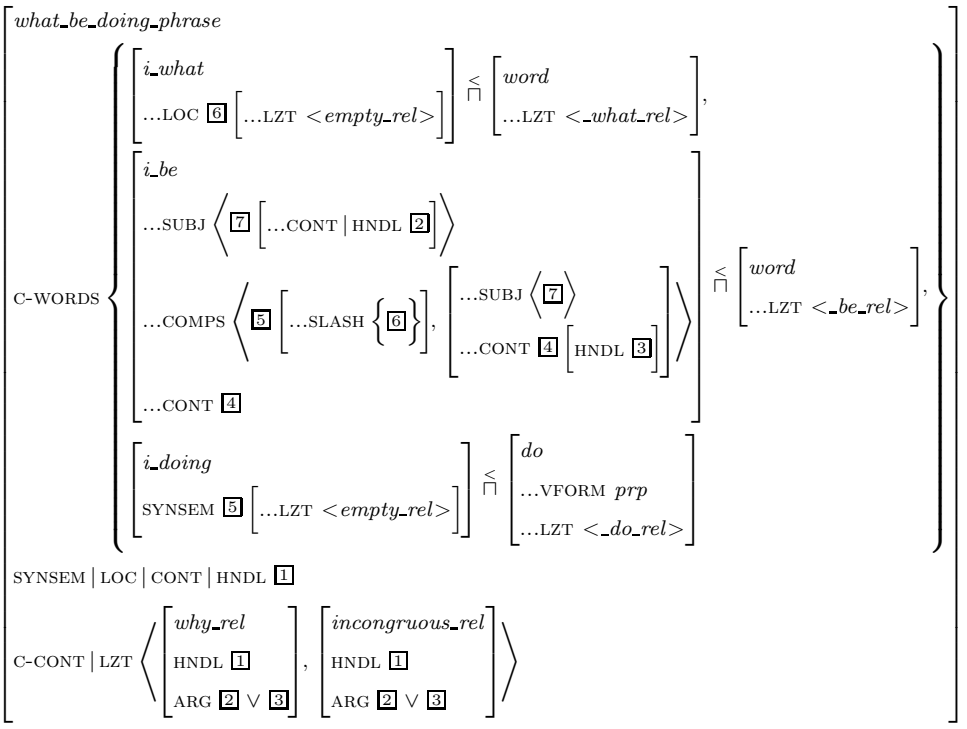
A plausible processing model should ensure that established compounds and canonical forms of idioms are always accessed first, because retrieval from memory is faster than productively deriving these forms. So in human language processing the second parse does not have to be considered. A computational model might be able to achieve the same effect using a mechanism similar to that needed for blocking and the 'elsewhere principle' (Zeevat 1995). However, this is one aspect of the approach proposed in this dissertation that requires further research, and I recognize that the idea of representing canonical forms of idioms in the grammar may be controversial. This is discussed in a bit more detail in Section 9.1.

5.3.10 Syntactic Constructions

The proposed approach can also be used for constructions like *what's the fly doing in my soup*, abbreviated WXDY for *What's X Doing Y* (Kay and Fillmore 1999), for example (247a). These can be described phrasally in spite of their syntactic flexibility (247b), because the syntax does not need to be fixed.

(247) a. What are your dirty feet doing on the breakfast table?

b. I don't know what Mary thought her feet were doing on the table.

(248)


In this representation for WXDY, the *what* and *doing* do not contribute to the meaning of the construction. Instead, the construction meaning of WXDY is roughly: why is X Y, and it is incongruous that X is Y.[17] E.g. the meaning of (248a) is:

(249) a. Why are your dirty feet on the breakfast table? It is incongruous that your dirty feet are on the breakfast table!

The construction meaning does not have to be localized on one of the words that make up the construction and there are no problems with scope. For example, if this construction were lexicalized on *be* as was suggested in an earlier version of Kay and Fillmore (1999), the 'incongruous' aspect of the meaning would have to be located there, and the question meaning would take scope over it unless this can somehow be prevented.

[17]The reason for the disjunction between 2 (the meaning of the subject) and 3 (the meaning of the complement) in the ARG of *why_rel* and *incongruous_rel* is that the representation has to be consistent with a quantifier in the subject or in the complement taking wide scope. This is a stand-in for the proper MRS treatment, which would involve the features TOP and LTOP, and qeq constraints (equality modulo quantifiers).

5.4 Alternative Variants of the Approach

There are several alternative ways one might go about developing a phrasal underspecification approach. They vary according to the precise way in which to refer to the words in a phrase, and in the mechanism used to relate idiomatic words to literal lexical entries. All of these approaches have some disadvantages compared to the approach presented in Section 5.2, and are therefore less desirable.

5.4.1 Variant 1

The simplest variant, sketched by Copestake (1994), is to constrain idioms to include other signs for which there are separate lexical entries. These should be related in some way to the lexical entries for the literal uses of these words, because at least the inflectional information is shared. This was done by default inheritance.

(250)
$$\left[\begin{array}{l} \textit{spill\_beans\_idiom\_phrase} \\ \textsc{words} \left\{ \left[\begin{array}{l} \textit{i\_spill} \\ \ldots\textsc{lzt} \left\langle \left[\begin{array}{l} \textit{i\_spill\_rel} \\ \textsc{und} \boxed{1} \end{array}\right] \right\rangle \end{array}\right], \left[\begin{array}{l} \textit{i\_bean} \\ \ldots\textsc{lzt} \left\langle \left[\begin{array}{l} \textit{i\_bean\_rel} \\ \textsc{inst} \boxed{1} \end{array}\right] \right\rangle \end{array}\right] \right\} \end{array}\right]$$

This approach presupposes the existence of independent lexical entries for the idiomatic words, which leads to the problem discussed in Section 5.2.1 of how to constrain their occurrence outside of the idiom.

It is also interesting to note that even if these independent lexical entries exist, it is necessary to complicate the phrasal constraints like (250) when stating certain kinds of restrictions on the flexibility of idioms, such as the inability of some idioms to occur passivized. For example, it is not sufficient to state constraints on the COMPS attribute in the lexical entry for the idiomatic verb, because this does not prevent the verb from undergoing a lexical rule like passivization. The same is true for specifying the VFORM of the verb. So these constraints have to be stated in the phrasal pattern as discussed above, or the lexical entry has to be specially marked to prevent this lexical rule from applying.

A further drawback of this approach is that there is no single place where the metaphorical mapping is located. Even though the idiomatic lexical entries may be related to the literal ones by default inheritance, this captures only part of the metaphor, because it is not done in the context of the other idiomatic words. For example, one can express that *beans* is related to *secrets*, but not that this is the case only in the context of *spill*. This piece of information is something that one can infer when looking at the grammar as a whole, but it is not directly expressed in one place that the mapping is between *spilling the beans* and *revealing the secrets*, which is in the family of *mind as a container* metaphors.

Apart from these differences this approach is similar to the one proposed in this dissertation and was the main inspiration for it.

5.4.2 Variant 2

In this approach only the idiomatic senses are mentioned in the LZT of the phrase, along with the relationships that hold between them. This is another suggestion made by Copestake (1994).

(251)
$$\left[\begin{array}{l} \textit{spill\_beans\_idiom\_phrase} \\ \\ \text{SYNSEM}\,|\,\text{LOCAL}\,|\,\text{CONT}\,|\,\text{LZT}\ \left\langle \ldots, \left[\begin{array}{l}\textit{i\_spill\_rel}\\ \text{UND}\quad \boxed{1}\end{array}\right], \left[\begin{array}{l}\textit{i\_bean\_rel}\\ \text{INST}\quad \boxed{1}\end{array}\right], \ldots \right\rangle \end{array}\right]$$

The idiom is represented as an idiomatic phrase which contains somewhere in its semantics two particular idiomatic relations (*i_spill_rel* and *i_bean_rel*), and therefore must contain the words *i_spill* and *i_bean*, because these are the only words with the right semantics.

This approach cannot handle non-decomposable idioms like *kick the bucket*, assuming that in the right analysis for such idioms there are no separate *rel*s to identify the words which occur in this idiom, and that there is a single *i_kick_bucket_rel* instead. If *bucket* has an *empty_rel* instead of a real meaning, then in this approach it cannot be distinguished from other idiomatic words without a meaning.

Note also that in this approach it is not possible to state syntactic constraints via the valence features as described in Section (5.3.9). Because there is no WORDS

feature there is no way to access the words in the leaf nodes of the syntactic tree except via the DTRs features, which as we saw in Chapter 4 does not allow for many kinds of variability.

Furthermore, this approach is like the previous one in that it presupposes the existence of lexical entries for the idiomatic words, and does not provide a place for the metaphorical mapping.

5.4.3 Variant 3

A third possibility assumes that the WORDS set does not include the entire literal lexical entries, but only their morphological paradigms. Because the semantic relations do not need to be changed, no defaults are needed. No separate lexical entries for idiomatic senses are needed. This approach shares some characteristics with the approach of Jackendoff (1997).

(252)
$$\left[\begin{array}{l} \textit{spill\_beans\_idiom\_phrase} \\ \text{WORDS} \quad \left\{ \left[\begin{array}{l} \textit{i\_spill} \\ \text{MORPH} \quad \textit{spill\_morph} \\ \ldots\text{LZT} \left\langle \left[\begin{array}{l} \textit{i\_spill\_rel} \\ \text{UND} \quad \boxed{1} \end{array}\right] \right\rangle \end{array}\right], \left[\begin{array}{l} \textit{i\_bean} \\ \text{MORPH} \quad \textit{bean\_morph} \\ \ldots\text{LZT} \left\langle \left[\begin{array}{l} \textit{i\_bean\_rel} \\ \text{INST} \quad \boxed{1} \end{array}\right] \right\rangle \end{array}\right] \ldots \right\} \end{array}\right]$$

In this approach representations for non-decomposable idioms would look like (253).

(253)
$$\begin{bmatrix} \textit{kick\_bucket\_idiom\_phrase} \\ \text{WORDS} \left\{ \begin{bmatrix} \textit{i\_kick} \\ \text{MORPH } \textit{kick\_morph} \\ \ldots\text{SUBJ} \left\langle [\ldots\text{INDEX } \boxed{1}] \right\rangle \\ \ldots\text{COMPS} \left\langle [\ldots\text{INDEX } \boxed{2}] \right\rangle \\ \ldots\text{LZT } <\textit{empty\_rel}> \end{bmatrix}, \right. \\ \begin{bmatrix} \textit{i\_the} \\ \text{MORPH } \textit{the\_morph} \\ \ldots\text{LZT } <\boxed{3}\ \textit{empty\_rel}> \end{bmatrix}, \\ \left. \begin{bmatrix} \textit{i\_bucket} \\ \text{MORPH } \textit{bucket\_morph} \\ \ldots\text{SPR} \left\langle \left[\ldots\text{LZT } < \boxed{3} > \right] \right\rangle \\ \ldots\text{INDEX } \boxed{2} \\ \ldots\text{LZT } <\textit{empty\_rel}> \end{bmatrix}, \ldots \right\} \\ \text{C-CONT} \mid \text{LZT} \left\langle \begin{bmatrix} \textit{i\_kick\_bucket\_rel} \\ \text{ACT } \boxed{1} \end{bmatrix} \right\rangle \end{bmatrix}$$

This third approach has several disadvantages. The fact that the morphology is the same is merely stipulated, and it is a coincidence that syntactic information is often shared between idiomatic occurrences of words and their literal counterparts. There is no location for the metaphorical mapping, and the literal entries are not available to describe idiom families.

5.5 Summary

As we have seen, an approach of the 'underspecified phrasal semantics' type is needed to handle all the properties of the data. A particular instance of this approach was outlined which compares favorably to various alternative variants.

The approach accommodates both decomposable and non-decomposable idioms, allows for the variability that idioms exhibit and the variety of types of information that can be fixed, and is crucially needed for the McCawley data. It also has the

advantage of not requiring separate lexical entries for the idiomatic senses of words. It is consistent with the psycholinguistic evidence available. Furthermore, phrasal hierarchies as used in this approach are independently motivated (Sag 1997), and many constraints are shared between idiomatic and non-idiomatic constructions. It is also possible for collocations to be listed in a similar way, which is necessary to model human sentence processing, and to speed up parsing in a computational system. The approach presupposes a 'flat' semantic representation of the kind provided by MRS. This can be seen as a challenge for semantic frameworks in which semantic nesting reflects syntactic structure.

Chapter 6

The Interaction of Idioms and Constructions

6.1 Introduction

In this chapter the interaction of idioms and constructions is investigated, in particular predicative idioms and the absolute construction.[1] The partially restricted distribution of some idioms provides further evidence for the primary ontological status of syntactic constructions. Specifically, the *with* and *with*-less absolute constructions (Stump 1985, McCawley 1983) are studied.

In the remainder of this chapter, Section 6.2 presents the data, Section 6.3 and Section 6.4 make the case for a constructional analysis of the data, Section 6.5 presents the constructional analysis, and Section 6.6 applies this analysis to the more nuanced data of a range of individual grammars.

6.2 Interaction Data

The particular data considered in this chapter are presented in (254) and (255). (254) gives examples of each of four predicative idioms (in italics) in the *with* absolute construction. *With* absolutes are sentence modifiers. This chapter is concerned with

[1]This chapter is based on work with Emily Bender (Riehemann and Bender 1999).

the case where they occur sentence-initially. Internally, *with* absolutes consist of the lexical item *with* followed by a small clause of the form NP + predicative XP.[2] In the examples in (254), the predicative idioms head the predicate of the small clause.

(254) a. With the negotiators still *poles apart* on so many issues, it's hard to see how these talks will ever end.
b. With expectations *flying high*, the Bulls have to win the championship this time.
c. With the media *all ears*, Clinton was very careful about what he said.
d. With peace talks *old hat*, it's hard to get a sense of hopefulness in the Middle East these days.

The examples in (255) are parallel to the examples in (254) except that they involve the *with*-less absolute construction.

(255) a. The negotiators still *poles apart* on so many issues, it's hard to see how these talks will ever end.
b. Expectations *flying high*, the Bulls have to win the championship this time.
c.?The media *all ears*, Clinton was very careful about what he said.
d.*Peace talks *old hat*, it's hard to get a sense of hopefulness in the Middle East these days.

In contrast to the *with* absolute examples, not all of the idioms are acceptable in the *with*-less absolute construction. This pattern of judgments is summarized in Table 6.1.[3]

The following two sections argue that these data motivate a constructional analysis. Section 6.3 first makes the argument that the absolute construction must be analyzed as a construction. Section 6.4 provides (further) motivation for a constructional analysis of the idioms.

[2]Here the class of predicatives is taken to be those phrases that can follow *be*, with the understanding that this distributional definition should line up with a cross-linguistically applicable semantic/pragmatic notion.

[3]There is a considerable degree of variation in this domain. Section 6.6 describes how the pattern shown in Table 6.1 was derived from the judgments of 14 speakers.

	with absolute construction	*with*-less absolute construction
poles apart	ok	ok
flying high	ok	ok
all ears	ok	?
old hat	ok	*

Table 6.1: Contrast Patterns Based on 14 Speakers.

6.3 Absolute Constructions

This section focuses on the *with*-less absolute construction, which has no overt lexical content uniquely associated with it. Nonetheless, it is a pairing of form (a small clause used as a sentence modifier) and meaning (its semantic and pragmatic properties). There are two possible ways to capture this pairing: a grammatical construction or a null element of some sort. Hantson (1992) develops a null complementizer analysis of *with*-less absolutes. Here, the null element $\emptyset$ in (256) is syntactically parallel to *with*, which Hantson takes to be a complementizer.[4]

(256) There he sat, $[_{S'}\ \emptyset\ [_{S}$ his back against the hot stones of the tower.]]

In general, there is a certain equivalence between null elements and constructions. It is difficult to imagine a paradigm that could be described with one but not the other, as long as one is solely interested in generating the right strings. However, approaches based on null elements and those based on constructions do differ in the kinds of generalizations they can capture elegantly. Here it will be argued that the distribution of predicative idioms across the two types of absolute constructions makes it possible to distinguish between the two approaches.

All of the idioms are acceptable in the *with* absolute, while only some are acceptable in the *with*-less absolute. On the null complementizer analysis, these data would have to be handled in terms of subcategorization of the null complementizer.

[4]Example (256) is adapted from Hantson (1992:89). This particular example is of a clause-final absolute. Hantson also discusses clause-initial absolutes and in fact draws no distinction, but does not happen to supply any examples of clause-initial *with*-less absolutes.

On the constructional analysis, they can be handled in terms of subtyping of the constructions (elaborated in Section 6.5 below).

Subcategorization is implausible because no other complementizers (or elements that select for clauses) are selective about lexical material in those clauses. Here lexical is used as opposed to grammatical—a complementizer could indeed select for clauses with a certain mood, where the mood is expressed on the verb. However, this is different from selecting for specific open class words. Idioms seem to be more like open class words in this respect than they are like grammatical properties such as mood.

On the other hand, Nunberg et al. (1994:516) mention several other idioms which are selective about which constructions they co-occur with. For example, idioms such as *Is the Pope Catholic?* only occur as questions and idioms such as *Break a leg!* only occur as imperatives. Thus while clause-selecting heads never care about the lexical content of those clauses, idioms are known to be selective about their context.

Further, an analysis in terms of subtyping makes it possible to specify exactly where these restricted idioms appear—instead of having to say where they do not. For every place one might expect an idiom to occur and it does not, the null complementizer analysis requires finding a selecting head and fixing its subcategorization so that it excludes that idiom. On a constructional analysis, the subtyping mechanism only requires dealing with the contexts in which the idiom does occur.[5]

However, if it were the case that the acceptability of these idioms turned on some semantic feature, then an analysis in terms of subcategorization (of a null complementizer or of a construction) would be more appealing. To test this possibility, a second survey was conducted with a separate group of 19 native speakers. These speakers were presented with the sentences in (254) and (255) plus parallel sentences with paraphrases for the idioms. The paraphrase sentences for *poles apart*, *flying high*, and *all ears* are as in (257).

(257) a. (With) the negotiators still *far apart* on so many issues, it's hard to see how these talks will ever end.

[5]Subtyping is not the only mechanism available in this approach. For an idiom that has the general distribution of a VP, for example, it would suffice to give it features like those of a VP.

b. (With) the Lakers *so successful*, LA fans are optimistic about the playoffs.

c. (With) the media *intensely alert*, Clinton was very careful about what he said.

The idiom *old hat* was harder to find a paraphrase for, perhaps because it may be undergoing a change in meaning. (Merriam-Webster's Collegiate Dictionary gives two meanings, 'old-fashioned' and 'lacking in freshness: trite', but a search of the North American News Text Corpus search turned up many examples which were not consistent with either.) A preliminary survey indicated that a meaning like 'commonplace' was the most current. The test sentences were changed to ones in which that meaning would be plausible, and participants were asked at the end of the survey what they thought the best paraphrase was.[6] The test sentences used for *old hat* were:

(258) a. (With) email and webbrowsers *old hat*, it's hard to remember what life was like before the Internet.

b. (With) email and webbrowsers *commonplace*, it's hard to remember what life was like before the Internet.

The results of this second survey were as follows. First, *commonplace* was judged to be the best paraphrase of *old hat* in these sentences by most of the participants and at least a possible paraphrase by most of the rest.[7] Secondly, there were 12 speakers who did not uniformly reject the *with*-less absolute. Of these, 8 accepted *old hat* in the *with* absolute but preferred *commonplace* to *old hat* in the *with*-less absolute. None of the remaining 4 speakers preferred *old hat* to *commonplace* in the *with*-less absolute.

Further, there were similar results for the idiom *all ears*. Because this idiom was generally more acceptable in the absolutes than *old hat* the numbers are smaller, but 5 speakers (of the 12) preferred *intensely alert* to *all ears* in the *with*-less absolute. Only one of the remaining 7 speakers preferred *all ears* in the *with*-less absolute.

[6]The choices offered were *boring, outdated, old fashioned, nothing new, nothing extraordinary, commonplace,* and *thoroughly familiar*. Another question was whether there was some better paraphrase not on the list.

[7]14 of the 19 speakers chose *commonplace* as the best paraphrase (possibly tied with some others) or said it was an acceptable one if they picked something else. 10 of the 12 speakers discussed below are in this category.

In all, only one speaker showed a pattern which would be consistent with a semantic subcategorization explanation for *old hat*. That is, he accepted the *with*-less absolute for other idioms and accepted both *old hat* and *commonplace* in the *with* absolute, but rejected both *old hat* and *commonplace* in the *with*-less absolute. However, this speaker was one of the 5 to get a contrast between *all ears* and *intensely alert*.

If the chosen paraphrases actually shared the relevant semantic properties of the idioms, then these results can be taken to show that the restricted distribution of predicative idioms across the two types of absolutes is not a matter of the two absolute constructions having different semantic restrictions. It can be concluded that the null complementizer approach, which relies on subcategorization, does not provide a natural account of the pattern of grammaticality shown in Table 6.1. In the next section an alternative, construction based account is discussed.

6.4 Predicative Idioms

The subtyping approach advocated (and described in section 6.5) requires a constructional approach to the idioms involved. This adds another piece of evidence arguing for a constructional approach to idioms to the many other motivations discussed in Chapter 2.

The distribution of these restricted idioms in a corpus across constructional and lexically selected contexts shows that there is constructional licensing of these idioms. The term 'constructional' is used where the licenser appears to be a construction (such as the *with*-less absolute construction) rather than a lexical item (such as *is* in *That's old hat*).

If constructionally licensed occurrences of idioms were rare, one might be tempted to write them off as peripheral, i.e., as instances where for whatever reason of performance or telegraphic register the licensing lexical item was omitted. However, constructional occurrences of idioms are not rare—in the corpus 22% of the occurrences of the four idioms do not involve a form of the verb *be*. Some of these involve other verbs like *seem*, but 9% of the overall occurrences are constructional. Further,

since the data is written and edited text, these instances cannot be explained away as slips of the tongue. Also, constructional occurrences are not restricted to telegraphic contexts such as headlines. This is too much data to be dismissed as 'peripheral', and it is not the type of data that could be attributed to performance factors. In fact, 'periphery' is a suspect concept because there is no principled way of determining where to draw the line between 'core' and 'periphery' (Kay and Fillmore 1999, Bender and Flickinger 1999).

Note that not all constructional representations of idioms are complete syntactic trees. Many idioms can occur discontinuously, as in (259).

(259) My music career was not *flying* all that *high* and I was tired of being on the road.

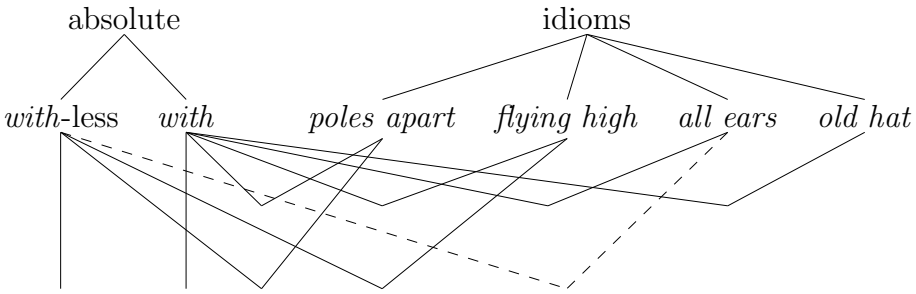
The corpus contains 14 occurrences of non-contiguous *flying high*, e.g., *flying* ***so*** *high (that)*. Other idioms are even more syntactically flexible, as was shown in Chapter 3.

So the distribution of predicative idioms is most satisfactorily analyzed in terms of a construction-based subtyping analysis. This provides further motivation for the constructional approach to idioms developed in this dissertation.

6.5 A Constructional Analysis

An illustration of how the formal device of subtyping captures the restricted distribution of these idioms is given in the partial type hierarchy in (260). Each node in the hierarchy is a construction type. The solid lines connect actually existing types. Only the types at the bottom of the hierarchy license grammatical sentences. Dashed lines indicate marginal types. (The dashed lines are not part of the HPSG formalism but rather a placeholder for a theory of marginality judgments.)

(260)



So, for example, this hierarchy expresses that the idiom *poles apart* can occur in the *with* and *with*-less absolute constructions, while the idiom *old hat* only has a mutual subtype with the *with* absolute construction—and therefore cannot occur in the *with*-less absolute. The closed-world assumption is used here. This means that only the types actually declared in the hierarchy exist. Since no common subtype for *old hat* and the *with*-less absolute is declared, that combination is not licensed. Note that the hierarchy displayed here is partial. All of the idioms have other subtypes, such as the one that allows *old hat* to occur with forms of *be*, which are not shown for space reasons. The absolute constructions each also have an unrestricted subtype. This subtype is the most common one for each construction. It combines freely with syntactically compatible phrases in the sense that those phrases can unify with one of its daughters, as is shown below. The idiom *old hat* cannot combine with either absolute construction in this way, because it has no subtype that includes just the phrase *old hat*.

The feature structure in (261) gives a partial description of the construction that licenses the *old hat* idiom.

(261)
$$\left[\begin{array}{l} \textit{old\_hat\_idiom\_ph} \\ \text{WORDS} \left\{ \begin{array}{l} \left[\begin{array}{l} \textit{i\_old} \\ \ldots\text{LZT} \ <\textit{i\_old\_rel}> \\ \ldots\text{MOD} \ \boxed{1} \end{array}\right] \overset{\leq}{\sqcap} \left[\begin{array}{l} \textit{word} \\ \ldots\text{LZT} \ <\textit{old\_rel}> \end{array}\right], \\ \left[\begin{array}{ll} \textit{i\_hat} & \\ \text{SYNSEM} \ \boxed{1} & \left[\begin{array}{l} \ldots\text{LZT} \ <\textit{i\_hat\_rel}> \\ \ldots\text{SPR} \ < \ > \\ \ldots\text{SUBJ} \ <\text{NP}> \end{array}\right] \end{array}\right] \overset{\leq}{\sqcap} \left[\begin{array}{l} \textit{word} \\ \ldots\text{LZT} \ <\textit{hat\_rel}> \end{array}\right], \ \ldots \end{array} \right\} \end{array}\right]$$

Here both *hat* and *old* are analyzed as semantically idiomatic. This may not be the right interpretation for all speakers, in particular those for whom the idiom means *old fashioned*. One could easily represent *old* as literal. The meaning of *hat* is changed to *i_hat_rel*, and it has the syntactic property of taking an NP subject instead of a specifier, i.e. it acts like other predicative nouns.

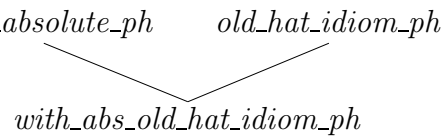
The feature structure in (262) describes the *with* absolute construction. It is a subtype of the absolute construction and, from it, inherits semantic and pragmatic information. More precisely, it inherits all the information it shares with the *with*-less absolute construction. This construction specifies constituent structure by means of its DTR (daughter) features. The HEAD-DTR is the lexical item *with* and the COMP-DTR is a predicative small clause.[8]

(262)
$$\left[\begin{array}{l} \textit{with\_absolute\_ph} \\ \text{HEAD-DTR} \ [\textit{with}] \\ \text{COMP-DTRS} \left\langle \left[\begin{array}{l} \text{PRED} \ + \\ \text{SUBJ} \ < \ > \end{array}\right] \right\rangle \end{array}\right]$$

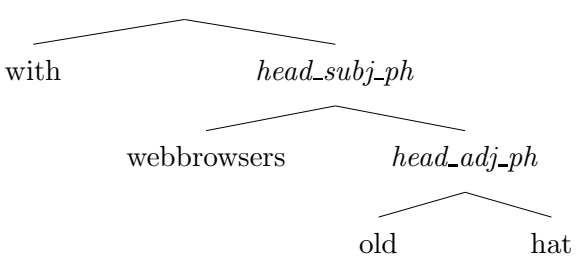
The feature structure in (263) is the subtype of the *with*-absolute construction (262) and the idiom *old hat* (261), as can be seen in (264).

[8]There have been proposals to replace these daughter attributes with relational constraints. The approach in this dissertation is consistent with that alternative representation for constituent structure. However, the analysis in this chapter is not consistent with a system in which *constructions* have a DAUGHTERS attribute while that attribute is not appropriate for *phrases*, because the research in this chapter suggests that it is necessary to have types that inherit from both constructions and phrases, so they need to be compatible.

(263)
$$\begin{bmatrix} \textit{with\_abs\_old\_hat\_idiom\_ph} \\ \textsc{comp-dtrs} \left\langle [\textsc{key}\ \textit{i\_hat\_rel}] \right\rangle \end{bmatrix}$$

(264)


Since the type *old_hat_idiom_ph* phrase does not specify any constituent structure, and since all nodes in the parse tree which dominate both *old* and *hat* are consistent with the constraints on *old_hat_idiom_ph*, it is not immediately obvious which node(s) in the tree are of this type. In fact, there is only one, as shown in (265). This is the only node which is compatible with all of the constraints on one of the subtypes of *old_hat_idiom_ph*.

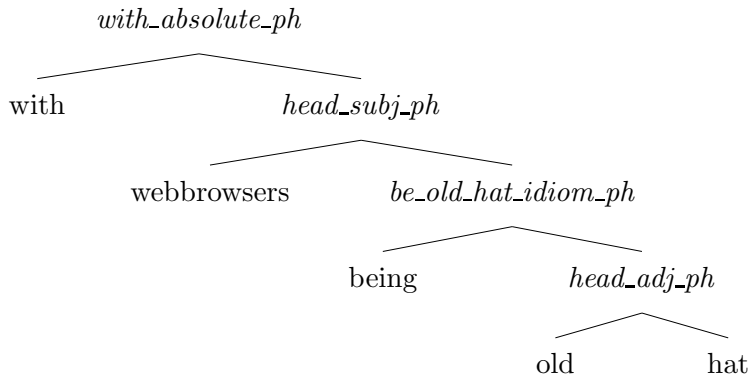
(265)


Note that one cannot specify on *with_abs_old_hat_idiom_ph* that the COMP-DTR is of type *old_hat_idiom_ph*. This is because in order to have that *old_hat_idiom_ph* be just *old hat* and not *be old hat* or *with* NP *old hat*, there would have to be a stand-alone subtype of *old_hat_idiom_ph*. This would nullify all of the advantages of the subtyping analysis.

The type *old_hat_idiom_ph* has another subtype, *be_old_hat_idiom_ph*. Phrases licensed by this subtype are free to combine with absolute constructions as part of the complement daughter. In this case, only the node labeled *be_old_hat_idiom_ph* in (266) is an instance of the type *old_hat_idiom_ph*.[9]

[9]At first glance, it looks like the node labeled *with_absolute_ph* might be compatible with the constraints on *with_abs_old_hat_idiom_ph*. But this is not the case, because in the example in (266) the KEY of the COMP-DTR is contributed by *being* rather than by (idiomatic) *hat*. Note that the fact that the type *old_had_idiom_ph* cannot occur at the root level makes the analysis in this chapter incompatible with the logically precise version of the approach in Chapter 5. But it is no problem for the intuitive version of the approach or its implementation in the style of Copestake (1993).

(266)



Unlike in (266), in (265) the *old_hat* idiom is involved at the same level in the tree as the *with* absolute construction. Therefore it is necessary to specify that the words *old hat* show up as the predicate of the small clause inside this construction, and not any further down the tree. That is, the matrix *with* absolute should not license *old hat* in (267).

(267)*With John thinking, "This issue old hat, I'd better move on," things are certainly going to get worse.

On the other hand, it would not do to say that the predicate of the small clause is exactly the two words *old hat*, since some modification is allowed:

(268) With these issues already old hat, we'll have to look for some more topics for position papers.

The solution is to have the type *with_old_hat_idiom_ph* specify that the primary semantic contribution of the complement daughter is that of the idiom *old hat*, as in (263). This works because the feature KEY always points to the semantic contribution of the syntactic head of a phrase. The KEY feature is shared between the head daughter and the mother in all headed constructions. In (268), *old hat* is the head of the head-modifier phrase *already old hat*, so the KEY gets passed up from *old hat*. In (267), the small-clause in the absolute is *John thinking ... move on*, which is headed by *thinking*, so it has a different value for KEY.

To summarize, here is how this analysis captures the important properties of these data. Both the absolute constructions and the idioms are represented phrasally. This has all the advantages discussed above, including naturally capturing the fact that

idiomatic words occur only as part of the idiom and cannot have those idiomatic meanings when they occur alone.

Given a phrasal representation for both the idioms and the constructions, the distribution of the idioms can be restricted by only allowing them to occur in certain environments. This is formally expressed by cross-classifying them only with some constructions but not others.

6.6 Individual Systems

In the first survey, 21 native speakers (8 linguists and 13 non-linguists) were asked for their judgments on the sentences in (254) and (255) plus some filler sentences, presented in random order. 7 speakers uniformly rejected the *with*-less absolutes. 10 speakers uniformly accepted the *with* absolutes. 9 speakers neither uniformly rejected nor uniformly accepted either of these constructions. Even with this small number of idioms studied, only 2 speakers had uniform judgments for both of these constructions, and it is possible that they might not be as uniform with other idioms.

The data presented in Section 6.2 were based on the statistical mode of the patterns given for each idiom by those 14 informants who did not uniformly reject the *with*-less absolutes. For example, the contrast for *old hat* as presented in (254d) vs. (255d) was the most frequent pattern given. Not everyone had the same contrasts, but 19 speakers (=90%) allowed some combinations of these idioms and syntactic constructions, and not others. A flexible analysis like the constructional analysis proposed here is necessary for all these speakers, although it may not be identical to the one presented here.

Two examples of individual speakers' systems can be found in Table 6.2 and Table 6.3. The first speaker does not have the contrast for *old hat* that was discussed above—he finds this idiom ungrammatical in both constructions. But an analysis like the one presented is needed to capture the contrast he has for *poles apart* and *flying high*, since this speaker did accept the *with*-less absolute, just not with any of these idioms. It is also needed for *all ears* and *old hat* because these were accepted by this speaker in other constructions, for example as a pre-nominal modifier (269).

	with absolute construction	*with*-less absolute construction
poles apart	ok	*
flying high	ok	*
all ears	*	*
old hat	*	*

Table 6.2: Contrast Pattern for Speaker 1

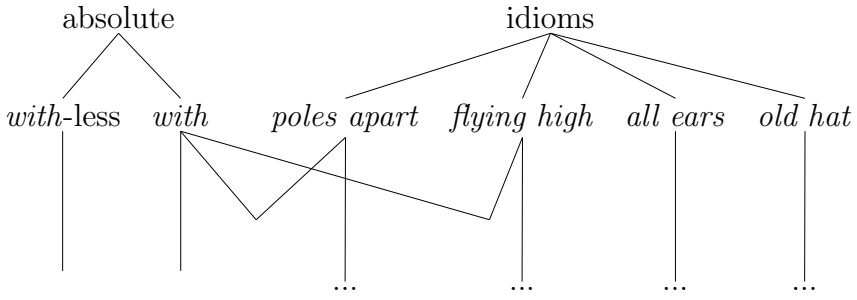
	with absolute construction	*with*-less absolute construction
poles apart	ok	ok
flying high	ok	ok
all ears	ok	ok
old hat	ok	*

Table 6.3: Contrast Pattern for Speaker 2

(269) We'll just get more old hat conspiracy theory journalism.

The hierarchy representing this speaker's system would be as in (270).[10]

(270)



The second speaker shows basically the pattern discussed before.

[10]The '...' in this hierarchy indicate that there are other subtypes of these idioms in addition to those involving the absolutes.

6.7 Summary

It was shown that the distribution of predicative idioms provides further motivation for a formal notion of construction. This approach handles the unpredictable distribution of these idioms by specifying the environments in which they can occur without missing the generalization that the SAME idiom occurs in all these contexts.

Note that it is not claimed that all idioms have to be treated this way. Some idioms may be totally permissive about the constructions they occur with, and for others, the restrictions might be explained semantically. It is also possible that there are some speakers who are totally uniform in their acceptance or rejection of idioms occurring in these absolute constructions. However, the study in this chapter showed that for at least 90% of the speakers some idioms require the kind of flexibility available in the approach described here. This is not surprising, since psycholinguistic evidence shows that speakers process canonical forms of idioms faster (McGlone et al. 1994). This suggests that speakers have representations for specific combinations of idioms and constructions, in addition to general knowledge of the idioms. It is an interesting question for further research how these constraints concerning relatively low-frequency data are learned.

Chapter 7

Derivational Morphology

This chapter presents a constructional approach to derivational morphology.[1] Approaches to morphology typically account for regular, completely productive affixation, while ignoring subregular and semiproductive schemata. The alternative approach to derivational morphology presented here relates exceptions and subregularities to productive rules. It accounts for the contribution lexicalized words make to the rule, and for the fact that not all new formations follow the 'rules'. It also captures linguistically relevant generalizations that cannot be expressed in other theories. The approach is formalized in terms of complex recursive schemata structured in a multiple inheritance hierarchy, without positing lexical rules or lexical entries for affixes. These schemata structure the existing lexicon, reducing redundancy, and at the same time serve as the basis for productive word formation. The approach handles zero-derivation and other nonconcatenative morphology straightforwardly.

7.1 Introduction

Morphological data are characterized by strong regularities as well as subregularities and exceptions. Because of the ubiquity of complex words with meanings that are not fully predictable from their parts, the matter of how these should be treated is important.

[1]The work in this chapter, except for the section on Hebrew, is based on Riehemann (1998).

In this chapter it will be argued that these cases should not be viewed as completely separate from the fully transparent words. If they were, linguistically significant generalizations would be lost, partial similarities between exceptional and non-exceptional patterns would not be expressed, and these words could not be seen as contributing to the determination of the rule.

The chapter concentrates on morphological derivation, as contrasted with inflection or compounding, and examples are mainly from German *bar*-adjectives (= English -*able*), which I analyzed in detail in previous work (Riehemann 1993). Further data come from *able*-adjectives in English, a less regular, more analogy based German suffix (-*ig*), and Hebrew instrument nouns.

It is traditional to suppose that the suffix -*bar* can be attached to the stems of all and only transitive verbs (e.g. *lesen* 'read' → *lesbar* 'readable'). This rule of -*bar* suffixation is assumed to be fully productive. Transitive verbs which do not allow -*bar* suffixation and non-transitive verbs that do allow such suffixation would constitute counterexamples to this rule. But such data do exist. Not only are there additional constraints on which verbs this suffix can be attached to, but there are lexicalized exceptions which do not conform to this productive rule. And most importantly, there are new formations which are not captured by the rule.

These data are a problem for most traditional approaches to morphology, namely those which assume *one* affix -*bar* and those which posit *one* lexical rule for forming *bar*-adjectives. These approaches have no mechanisms for dealing with the counterexamples. One might think it possible to refine the rule in such a way that it would describe all and only the correct cases. But the data below suggest that this cannot be done—there does not seem to be a single generalization that is general enough to encompass all existing and possible *bar*-adjectives and at the same time specific enough to exclude impossible examples. It might be possible to construct a complex and highly arbitrary 'rule' that lists all the lexicalized exceptions in its definition. But that would be very unintuitive, and furthermore would not be able to account for productive exceptions. Therefore I will continue to assume as it has been assumed in the literature that 'the rule' is that -*bar* attaches to transitive verbs, and develop an approach in which this regularity is expressed while the exceptions can be handled

as well.

In the constructional approach presented here, -*bar* is neither a suffix with its own lexical entry and strict subcategorization information, nor is it just phonological material added by lexical rule. Instead it is seen in terms of a schema, arising as a generalization about existing *bar*-adjectives in the lexicon. The schema says that there is a class of adjectives ending in -*bar* which have transitive verb stems as their first part. It also states various semantic and syntactic properties and relations to the verbal stem. Some of these adjectives are lexicalized subtypes of the pattern, but more can be formed productively according to the constraints embodied in the schema. The system allows for complex and recursive derivation because the transitive verb stems can be complex themselves, and because the resulting adjectives are of the appropriate type to undergo further derivation.

The main *bar*-pattern is seen as a generalization about existing words, just as a subregular one is. These patterns are hierarchically structured with respect to each other, which expresses their relationship. The approach not only accounts for more of the data, but also captures more generalizations, and could be the beginning of a theory of how regular affixations develop and are acquired.

The data will be treated using lexical multiple-inheritance hierarchies as used in HPSG. This framework is particularly well suited for expressing these generalizations because it is directly concerned with cross-classifying relations among linguistic types. Also, the rich lexical representations that are motivated by the data can be expressed clearly in HPSG, and fit well with its general assumptions.

Many linguists and psycholinguists, for example Jackendoff (1975), assume that many complex derived words are stored individually in the (mental) lexicon. This means that word formation differs from syntax in that frequent words are not derived 'on-the-fly' every time they are produced, but are stored as units even if they have an internal structure and are predictable from those parts. Evidence for this view comes e.g. from Rainer (1988) and Becker (1990), who observe that even frequent words which are the predictable output of a productive rule can block other words in some cases. For example in Italian, quality nouns for adjectives in -*ido* can be productively formed with the suffixes -*ezza* and -*ità*, and generally speakers accept both forms. But

when one of the forms is frequent, e.g. *avidità* ('greediness'), speakers do not accept the alternate form, even though the frequent form is not irregular in any way. So in principle lexicalization is independent of idiosyncrasy, although in most cases some unpredictable aspect is involved as well. Once a word is lexicalized it can undergo semantic drift, and acquire idiosyncrasies of all kinds, resulting in a meaning that is different from what would be expected if the word were formed productively and the meaning were completely compositional. So the phenomenon of semantic drift also shows that lexicalization has taken place.

The constructional approach presupposes this view because the existence of the entries for these lexicalized words, which correspond to the types at the bottom of the hierarchy, is necessary to explain which schemata for word-formation exist. But in the proposed approach only the idiosyncratic information about these complex words has to be listed, with the remaining information being filled in by inheritance. So unlike the analysis in Jackendoff (1975), this analysis does not require a separate evaluation measure for determining the contribution of lexical rules to reducing the information content of the lexicon. Another difference between Jackendoff's redundancy rules and the approach proposed here is that the latter is able to express not only relationships between words but also the relationships between rules. However, the approach is similar enough to Jackendoff's ideas to think of it as an incorporation of those insights into an HPSG architecture with the above-mentioned improvements.

Any adequate theory should provide a means of stating the idiosyncratic information in lexical entries, and at the same time it should make precise which parts are predictable and correspond to aspects of a productive rule. This is desirable on a descriptive level, where information would otherwise have to be repeated redundantly. While it could be argued that the partial regularities have historical reasons and do not need to be taken into account by synchronic grammar, that view becomes implausible when considering how large a number of 'regular' complex words need to be listed because of some slight specialization of meaning or other unpredictable properties. Approaches that do not relate lexicalized words to rules in any way also do not predict that existing words in the language will have an effect on what the rules of morphology are, and do not have anything to say about how they could be

acquired.

The strongest evidence for the relevance of the existing lexicon is that it is possible to form new words on the basis of existing 'subregular' patterns wherever it is sufficiently clear how the new word would relate to its stem, and how it would be interpreted. This kind of word formation according to patterns determined by a small set of existing words, and sometimes even by analogy to only one word, is needed for a full account of *bar*-derivation. It is widely acknowledged that such a process is the basis of some other kinds of word formation processes (see e.g. Motsch (1988) or Plank (1981) for more examples of semiproductive morphology). Since it is shown that even the highly regular *bar*-derivation cannot be described adequately by a single rule, it will be argued that such a rule is best seen as a special case of word formation on the basis of existing patterns, a very general regularity among existing words.

7.2 The Morphological Data

7.2.1 German bar-Adjectives

The *bar*-adjective data come mostly from various corpora of German. The main corpus consists of newspaper texts sampled between 1985 and 1988 (10 million words, 17,292 *bar*-adjective tokens, 836 types). I also looked at other newspaper and more general corpora, and the language used in contributions to electronic newsgroups. For details see Appendix A. This resulted in a total of 1226 different types of *bar*-adjectives, most of which occur only once or twice. Many of these low-frequency words seem to have been productively formed.[2]

The use of corpora was important because it provided the subregularities and exceptions that I probably would not have found through intuition alone. I also used native speakers' judgments to confirm the acceptability of the examples from the

[2]Of course I do not claim that all words that happen to occur only once in my corpora are productively formed, or even that it is possible to determine this for any particular example—what is productive for one speaker may be lexicalized for another. But most of these infrequent words are completely regular and predictable, and according to speakers' intuitions they are not established as lexicalized words of German. And on a more general level, affixes with large numbers of infrequent types are probably more productive than affixes with few infrequent types.

corpora[3] and the unacceptability of the examples needed for the argumentation.

The prototypical *bar*-adjective is derived from a transitive[4] verb. An example is *bemerkbar* 'noticeable':

(271) *Sie bemerken die Veränderung. Die Veränderung ist bemerkbar.*
They notice the change. The change is noticeable.

As in passives, the accusative object of the verb becomes the subject of the adjective. But the dropped subject normally cannot be expressed in *by*-phrases. Semantically a notion of possibility is added—something that is *lesbar* can be read.

It is traditionally assumed that the suffix *-bar* is fully productive and can attach to all transitive verbs. While it is true that it can attach to many transitive verbs to form new *bar*-adjectives, there seem to be additional semantic constraints.

Toman (1987) proposed that the verbs have to be (weakly) intentional, which includes not only verbs in which the subject is an intentionally acting agent but also those where at least 'an effort can be made'. As examples of the latter he gives *erkennen* ('recognize'), *bemerken* ('notice') and *erraten* ('guess'). This constraint is intended to exclude examples such as ?*verbitterbar* ('embitterable'), ?*enttäuschbar* ('disappointable') and ?*überraschbar* ('surprisable'). While it is disputable whether these really should be excluded, it is clear that restrictions in addition to transitivity are needed to exclude **wiegbar* ('weighable') or **dauerbar* ('lastable'), as in examples like (7.2):

(272) * *1 Kilogramm ist (von dem Buch) wiegbar.*
1 kilogram is (by the book) weighable.

But an absolute intentionality constraint would also wrongly exclude possible examples like *abbaubar* ('decomposable'), *resorbierbar* ('absorbable'), *regenerierbar* ('regenerable'), and *verformbar* ('deformable'):

[3]Examples from corpora need some verification since they may be typos, etc. But the occurrence in a corpus plus verification is much stronger evidence than mere acceptance by a native speaker. And researchers' intuitions may have been affected by studying a phenomenon for too long. Corpora can also provide the right data to present to one's informants, examples that might otherwise have escaped attention.

[4]'Transitive' is here and throughout to be understood as referring to verbs which have an object bearing accusative case.

(273) *biologisch leicht abbaubare Stoffe*
biologically easily decomposable substances
'easily biodegradable substances'

For more details about the nonmorphological properties of *bar*-adjectives see Riehemann (1993).

Still -*bar* affixation is a highly productive process. In the main corpus, more than half the types (435) occur only once or twice, accounting for only 3.15% of the tokens. Most of these are not listed in any dictionary, and people use them even if they do not remember having heard them before. This suggests that they have formed these words productively. This becomes even more obvious from the fact that the suffix can be attached to verbs that have newly entered the language—*faxbar* 'faxable' used to be such a word, and currently *ftpbar* 'ftpable' and *downloadbar* 'downloadable' are beginning to appear on a few web pages.

The corpora also show that there are a number of frequent, lexicalized *bar*-adjectives. In fact, 7% of the types account for 70% of the tokens. These words have exceptional properties of all kinds:

- phonological:
  - dropping of '-ig' in the stem
    (*entschuldigen* 'excuse' ⇒ *entschuldbar* 'excusable')
- semantic:
  - additional aspect of meaning
    (*essen* 'eat' ⇒ *eßbar* 'safely edible')
  - obligation instead of possibility
    (*zahlen* 'pay' ⇒ *zahlbar* 'payable' ('payable by the 15th' = 'has to be paid by the 15th'))
  - lexicalized in one particular sense
    (*halten* 'hold, keep' ⇒ *haltbar* 'non-perishable' ('keep-able'))
- syntactic:
  - from verbs with dative objects
    (*entrinnen* (+ Dat) 'escape' ⇒ *unentrinnbar* 'inescapable')

  - from verbs with prepositional objects
    (*verfügen über* 'have at one's disposal' ⇒ *verfügbar* 'available')
  - from reflexive verbs
    (*sich regenerieren* 'regenerate' ⇒ *regenerierbar* 'regenerable')
  - from intransitive verbs
    (*brennen* 'burn'[5]⇒ *brennbar* 'inflammable')
- highly exceptional:
  - no verbal stem (*sichtbar* 'visible')
  - no notion of possibility (*fruchtbar* 'fruitful')

These examples illustrate the well-known fact that complex words cannot always be interpreted in a completely compositional fashion, and that lexicalization can result in semantic drift and the development of idiosyncrasies of various kinds. But most of these words are also still clearly related to the 'rule', i.e. they have some properties in common with the completely regular *bar*-adjectives. For example, most of them share the notion of possibility in their semantics and the fact that the verb's object becomes the adjective's 'subject'. Listing them with full specifications would mean having to repeat predictable, transparent information.

Also—unexpectedly and more importantly—there are *new* formations which do not conform to the pattern of the fully 'regular' ones. The assumption is that infrequent words which one has not consciously heard before and which are not listed in dictionaries are too rare to have been learned as exceptions.

- from verbs with dative objects:
  - *unausweichbar* 'inescapable'
  - *unentfliehbar* 'unfleeable'
  - *unwiderstehbar* 'irresistible'
  - *vertraubar* 'trustable'
- from verbs with prepositional objects:
  - *quantifizierbar* 'quantifiable'

[5]Unlike English *burn*, German *brennen* cannot be used transitively.

  - *unzweifelbar* 'undoubtable'
  - *verzichtbar* 'abstainable'
  - *zugreifbar* 'accessible'

- from reflexive verbs:
  - *deformierbar* 'deformable'
  - *erneuerbar* 'renewable'
  - *verflüchtigbar* 'evaporatable'
  - *zersetzbar* 'decomposable'

- from intransitive verbs:
  - *unausbleibbar* 'inevitable'
  - *ungerinnbar* 'uncoagulatable'
  - *verrottbar* 'decayable'
  - *verheilbar* 'healable'

These adjectives are formed from verbs which do not have an object in accusative case[6], thereby violating what is seen in traditional word-syntactic approaches as the 'subcategorization requirements' of the affix.[7] But because these words do have parallels among the lexicalized subregularities, there are generalizations present in the constructional approach to morphology that serve to structure the existing lexicon. These can be used as a basis for semiproductive word formation.

7.2.2 English able-Adjectives

The same kinds of observations can be made for English *able*-adjectives, although it is much harder to find examples that are unquestionably derived from intransitive

[6]It is impossible to know whether the people who used these words shared the intuition that the reflexive verbs above do not have transitive counterparts from which the adjectives could have been formed. But it is likely that they do because there are no transitive uses like **die Natur erneuert diese Energiequelle* corresponding to examples like *erneuerbare Energiequellen* 'renewable sources of energy'.

[7]If exceptions could be explained semantically, then there would be no exceptions, and everything could be done by one rule. Thematic aspects certainly play a role in determining which of the exceptional words get produced, but it is not a hard and fast constraint, i.e. not everything that is a theme can be used. This can be seen for example with intransitive *fallen* 'fall' (**fallbar*), or when the object is in dative case (**helfbar* 'helpable').

verbs, since unlike in German it is often possible to use even the underlying verbs transitively (with a causative meaning). Examples of this kind are *detonable*, *lapsable*, *rottable*, *unwiltable*, or *burstable*. These are found in the corpora, but could arguably have been derived from the respective transitivized forms of the verbs.

The following *able*-adjectives, all of which are attested in corpora and/or on the web, were judged acceptable but not established words of English by the majority of the eight native speakers I asked, while the same speakers thought that their verbal stems could not be used transitively.[8]

- from intransitive verbs:
  decayable, *demurrable*, *expirable*, *materializable*, *mutatable*
- from verbs with prepositional objects:
  abstainable, *adherable*, *compliable*, *consentable*

These words are arguably too rare to have been learned as exceptions by native speakers. A Lexis-Nexis search of 'General News' from 'Major Newspapers' for 'all available dates' (more than 20 years) conducted on 5/28/01 found only 3 occurrences of *demurrable*, *mutatable*, and *compliable*, 1 occurrence of *abstainable*, 1 nominalized occurrence of *decayable*, and no occurrences of *expirable*, *materializable*, *adherable*, and *consentable*.[9] So these are not established words of English, but appear to have been productively formed, even though they do not conform to the 'rule' of -*able* affixation.

[8]The judgments about the substantially longer list of adjectives varied greatly among speakers, with some accepting almost none of the words, while most others thought almost all were acceptable and some even common. This is not surprising if word formation is indeed based on the existing lexicon in the way proposed here, since the possibility of forming (and accepting) new instances should depend on the number of words of this kind someone has come across previously, and whether a generalization has been formed.

[9]
- [...] the consequences of allowing a really *demurrable* claim to proceed imposed on the judge an obligation to look at it with particular care (The Times, 4/2/1998)
- This gene is recognized as the most commonly *mutatable* gene in human cancer (The San Diego Union-Tribune, 9/18/1992)
- It's 19th- and 20th-century impermanent art made of wood, paper, cloth, feathers, shells and other *decayables*. (The Washington Post, 12/2/1983)
- The Single *Abstainable* Vote (The Independent, 3/24/1992)
- For the photographers' pleasure, Ms Tottman showed off the dogs' abilities by putting the *compliable* Strapper around her neck, where it proceeded to give a passable impression of a fox fur. (The Times, 2/26/1999)

Even clearer evidence comes from recent computer terminology. I found several occurrences each of *grepable/greppable*, *lynxable*, and *telnetable/telnettable* on the web, while the corresponding verbs do not seem to be used transitively.[10] These words must have been productively formed in recent history by at least one individual, probably by several people independently.

7.2.3 German ig-Adjectives

German *ig*-adjectives are a good example of two things: how one affix can have various specialized subpatterns, and how the formation of new words follows these patterns. It is not sufficient to encode only the information they share at a higher level.

The suffix *-ig* can be used to form adjectives from a wide variety of categories: nouns (*Staub, staubig* 'Dust, dusty'), verbs (*zappeln, zappelig* 'fidget, fidgety'), adverbs (*sofort, sofortig* 'immediately, immediate'), and nonproductively, adjectives (*voll, völlig* 'full, fully') (Fleischer and Barz 1992).

But there are fine semantic restrictions even within the productive categories. For example the *ig*-adjectives formed from nouns all mean something like 'characterized by' (*Rost, rostig* 'rust, rusty'). But it is not possible to form new *ig*-adjectives from just any noun—for example *knopfig* (as in **knopfiger Anzug* 'buttony suit') is not a possible word (Motsch 1977), although it would seem to make sense semantically if a suit is characterized by having a large number of, or otherwise very noticeable, buttons.

This fact seems puzzling, but on further examination of the data it turns out that the existing *ig*-adjectives fall into various subgroups semantically. Some have a meaning of 'like (an) X' (*Kalk, kalkig* 'lime, limy'), others combine with nouns denoting concrete physical objects and mean 'characterized by the presence of substances (on the object)': *fleckig* 'spotted', *schmutzig* 'dirty', *ölig* 'oily', *dreckig* 'dirty', *staubig* 'dusty', *fettig* 'fatty'. Motsch argues that the meaning has to involve 'impurities', and it is true that new formations of this kind tend to have this kind of negative meaning (*Siff, siffig* 'filth, filthy'). The acceptability of new formations is inversely

[10]Some speakers say that *grep* can be used transitively. However, there are many speakers who would only say *to grep for a string in a file* but not *to grep a file*, yet use *the file is grepable*.

proportional to the distance from existing patterns. ??*Politurig* 'polishy' would be an example of 'characterized by the presence of non-negative substance (on object)', and does not seem acceptable. We have already seen an unacceptable example of a non-substance on an object (**knopfig*). And it is impossible to form new words just on the higher level meaning 'characterized by' that all the various patterns seem to have in common **pausige Konferenz* 'pausy conference', **buchiges Zimmer* 'booky room'.

New formations in analogy to the existing semantic groups, on the other hand, are common. *Deppig* 'foolish' is an example with a 'like' meaning, *fusselig* 'linty' one of the 'characterized by contaminating substance (on object)' type, *buggig* 'buggy' one of the 'characterized by the negative presence of'.[11]

This shows that *-ig* is not simply productive with nouns, but it is necessary to look at the data in more detail and determine what kinds of nouns it is found with, since the possibilities for new formations reflect the existing data.

7.2.4 Semi-Affixes

In the morphology of German, there are entities traditionally called 'Affixoide' that are said to have a status that is intermediate between an affix and part of a compound. Examples of these are *-voll* 'full', *-frei* 'free', *-arm* 'poor', *-zeug* 'stuff', *-wesen* 'being', and *-gut* 'property',[12] (Olsen 1988). Each of these is an existing word of German, and has a meaning related to that word when it occurs in a complex morphological structure. But these words have also developed specialized functions within such structures, which needs to be captured. Their meaning is usually more general and abstract than that of the corresponding simplex words. For example, *Zeug* means 'stuff' (usually pejoratively), and the words *Strickzeug* (stuff needed for knitting), *Schulzeug* (stuff needed for school), and *Putzzeug* (stuff needed for cleaning) are clearly related to that meaning, but do not have the pejorative aspect, and all share

[11] A Google search of German language documents on 1/19/01 found 8 occurrences of *buggig*, which were apparently written by 8 different people. Note that the English word 'buggy' is not unknown to German computer users, which probably played a role. However, the choice of German suffix was still in accordance with the existing semantic groups.

[12] These are glosses for the independent words.

a semantic 'purpose' relation.[13] Although this relation is frequently found in other compounds, too, it is especially relevant here, because the fact that there is such a group of words ending in *-zeug* with that meaning is the basis for forming new ones. So, even if one treats them as compounds formally (as Olsen suggests) it is necessary to pay attention to the existing lexical items and the patterns among them, since it is not sufficient to assume that the lexical entry for *Zeug* together with the rules of compounding will give the right results. One does not find compounds with *-zeug* involving a 'belong to' or 'part of' relation, for example.

7.3 Previous Approaches

In this section previous suggestions for dealing with derivation in theories of morphology are summarized briefly, and some of the problems they run into are pointed out.

Two major approaches to an account for the regular aspects of word formation have been suggested in the linguistic literature: one is based on processes, i.e. LEXICAL RULES, the other on the arrangement of items, i.e. a WORD-SYNTAX.

7.3.1 Lexical Rules

The first approach is exemplified for instance by Aronoff (1976) or Jackendoff (1975) (following Stanley (1967)), who assume lexical rules or lexical redundancy rules. Here the affix is introduced syncategorematically in the rule as in (274), representing information about the word's derivation. The first part is a verb, but the affix is added by the rule and is not itself marked for category.

(274) $[[\text{les}]_{\text{V}}\text{bar}]_{\text{A}}$

7.3.2 Word Syntax

The second approach could be called phrase-structural, and, for a complex word like *lesbar*, would posit the internal structure shown in (275):

[13]The word *Flugzeug* ('airplane') refers to a single object and is a lexicalized exception.

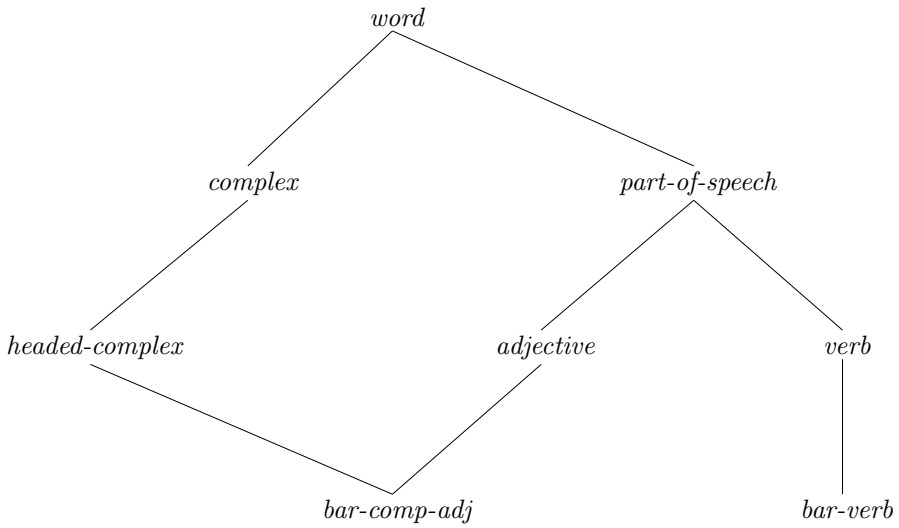
(275) $[[\text{les}]_{\text{V}}[\text{bar}]_{\text{A}^{\text{af}}}]_{\text{A}}$

Here the first part is the stem of the verb *lesen* ('read'), and the second the affix *-bar*, with a phrase structure rule allowing the two to combine to yield a complex adjective, where the category is determined by the suffix. An approach of this kind was proposed for example by Selkirk (1982). See Spencer (1991) for references to related work, and an overview of the item-and-arrangement tradition in morphology.

There has also been an implementation in HPSG of a word-syntactic approach. Krieger and Nerbonne (1993) and Krieger (1994) use a phrase structural approach to morphology to illustrate their inheritance lexicon. They assume that derivational affixes have category marked lexical entries and that they combine with other material, subject to various morphological principles.

Words are classified both according to their category and their morphological properties, as can be seen in (276), part of the hierarchy from Krieger and Nerbonne (1993).

(276)



Headed-complex words are taken to be subject to various principles, e.g. the Morphological Head Feature Principle (see below).

In Krieger and Nerbonne's approach *bar*-adjectives are represented as a word class under which the exceptional lexicalized instances are nonmonotonically grouped. This means that irregular forms can have properties that override the more general regular

information given in the class definition. The definition states that a *bar*-adjective has two parts, a *bar-suffix* and a *bar-verb*, as can be seen in (277).

(277)
$$\begin{bmatrix} \textit{bar-comp-adj} & \\ \text{MORPHS} & \begin{bmatrix} \text{HEAD-MORPH} & \textit{bar-suffix} \\ \text{COMP-MORPH} & \textit{bar-verb} \end{bmatrix} \end{bmatrix}$$

The lexical entry for -*bar* which Krieger (1994) assumes is given in somewhat abbreviated form in (278):

(278)
$$\begin{bmatrix} \textit{bar-suffix} & & \\ \text{MORPH} & \begin{bmatrix} \text{FORM} & \textit{bar} & \\ \text{SUBCAT} & \begin{bmatrix} \textit{bar-verb} & \\ \text{SYN}\,|\,\text{LOC}\,|\,\text{SUBCAT} & \begin{bmatrix} \text{OBJ} & \boxed{1} \\ \text{COMPS} & \boxed{2} \end{bmatrix} \\ \text{SEM} & \boxed{3} \end{bmatrix} & \end{bmatrix} & \\ \text{SYN}\,|\,\text{LOC} & \begin{bmatrix} \text{HEAD}\,|\,\text{MAJ} & A & \\ \text{SUBCAT} & \begin{bmatrix} \text{SUBJ} & \boxed{1} \\ \text{OBJ} & \textit{NIL} \\ \text{COMPS} & \boxed{2} \end{bmatrix} & \end{bmatrix} & \\ \text{SEM} & \begin{bmatrix} \text{OPERATOR} & \Diamond & \\ \text{SCOPE} & \boxed{3} & \end{bmatrix} & \end{bmatrix}$$

So the suffix -*bar* is represented as a lexical entry that has all the syntactic and semantic properties of an adjective. Because the suffix is the head of the resulting complex adjective, these properties become the adjective's properties via the Morphological Head Feature Principle. The suffix also has the morphological property of wanting to combine with a particular kind of verb, a *bar-verb*, and it integrates various properties of this verb with its own properties. For example, it equates its subject with the verb's object.

7.4 Problems of these Approaches

7.4.1 Affixes and Stems are Not Free

The fact that affixes cannot occur by themselves but only as parts of complete words is an integral part of the lexical rule approach, but not of the word syntactic approach, where additional stipulations are needed to make sure they phonologically attach to the stems they combine with.

The fact that stems can never occur by themselves needs to be stipulated in both approaches, and it is much harder to do in word syntactic approaches since they have to be 'free' to combine with some material (affixes) using basically the same mechanism that they have to be prevented from entering in otherwise. And it is not possible to restrict their occurrence through subcategorization. This is because affixes select for the stems they can occur with, but not vice versa.

7.4.2 Sub-Patterns Needed to Structure Lexicon

As we have seen there are many lexicalized words with internal structure and other properties shared with words of the same type. Redundancy can be avoided and generalizations captured by structuring the existing lexicon and organizing it hierarchically. It is possible to do this in addition to a word-syntactic or lexical-rule approach like Krieger and Nerbonne (1993). But this would not be an integrated view—the fact that these lexicalized patterns exist would not have anything to do with the fact that the rule or productive affix exists, and therefore would not be expected to play a role in how the rule developed historically or how it could be acquired.

7.4.3 Additional Mechanism Redundant

If one looks at the Krieger and Nerbonne (1993) approach closely, it turns out that having both this classificatory approach to morphology *and* a word-syntactic treatment at the same time leads to several redundancies. One of these is that the category is multiply determined—once by inheritance, since *bar-comp-adjective*s are a subtype

of adjectives, and once by the MORPHOLOGICAL HEAD FEATURE PRINCIPLE, which states that the category of the complex word is the same as that of the affix. Also, the information as to which verbs the affix can combine with is given twice: in the constraints on *bar-comp-adjective*, and also in the subcategorization requirements within the lexical entry of the affix. Finally, for lexicalized instances which are of type *bar-comp-adjective*, one still needs to make sure that the first part is appropriately marked as a *bar-verb*. All this indicates that the hierarchical approach alone, if properly extended, should be able to capture the facts.

7.4.4 Sub-Patterns Needed for Subregular Productivity

In both word syntactic and lexical rule approaches subregularities and exceptions are a problem, since in the first, a clear subcategorization requirement needs to be stated for the affix, and in the second, one has to state which class of words the rule applies to.

Furthermore, for less regular suffixes such as *-ig*, one would need 'lexical entries' for all the various subcases occurring with different noun classes. This would be a little less troublesome if they were organized into a hierarchy, but still it seems to be the case that the more complicated this gets, the more sense it makes to organize the *words* into the hierarchy, rather than the affixes. Such a hierarchy of words is needed in any case, for the lexicalized forms, and it would have to be duplicated in the affix hierarchy. Also, once it becomes necessary to refer to arbitrary semantic information in the 'subcategorization' information for the affix, it becomes much less similar to syntactic subcategorization, and it is questionable whether it should still be analyzed that way.

However, if lexical sub-patterns already exist, subregular productivity can be understood to be based on them. Given that there is a class of words having these semantic properties, one can say that new ones can be formed precisely because of the existence of that pattern.

7.5 Outline of the Proposed Approach

For all the reasons described in the previous section, it seems best to see affixes as generalizations about classes of words. What is proposed here is neither word syntax nor lexical rules, though closer to lexical rules, since it does not give the suffix an independent existence. One could of course think of an entire schema as being a unusual kind of 'lexical entry' for the suffix. But the proposed analysis has the advantage of generalizing readily to non-affixal morphology.

I assume an internal morphological structure for derived words similar to that in (274). But unlike in lexical rule approaches this structure is explicitly represented as part of the information about a complex word, similar to the coding of phrase structure, although in contrast to (275) the affix does not have its own lexical entry and is not marked for category.

In the proposed approach, there is a schema expressing the fact that there is a class of words, ending in the suffix -*bar*, that have transitive verbs as their morphological basis. It also states how the syntax and semantics of the verb relates to that of the adjective. For example, the accusative object of the verb is linked with the subject of the adjective, and the semantics of the verb reappears within the scope of the possibility operator in the semantics of the adjective.

Because these schemata contain this additional information, they are not just phonological like those in Bybee and Slobin (1982). These authors give evidence that children form product-oriented phonological schemata of complex words, and for example do not try to add an affix to a word that already appears to contain that affix. Only later do they 'learn the rules' of affixation.

Perhaps one can link up the two views. It is not implausible that children, when learning complex morphology, should first form phonological generalizations, since these are most directly available. The more complex schemata, as used in the proposed constructional approach, could then be formed later. Once children realize that the class is not merely defined by its characteristic ending, but also that the first part has to be a stem of a certain kind, they can go on to discover the systematic syntactic and semantic relationships of the complex words to those stems.

This may be part of the explanation of how 'rules' are acquired. They could start out as generalizations about rote-learned examples and gradually become more complex, while remaining of the same kind and never losing their connection to existing words of the language.

In the proposed approach there are also generalizations for the other (subregular) patterns. These analyze far fewer existing words, which is part of the explanation of why they cannot be used unconditionally in productive word formation.[14]

All generalizations serve primarily to organize the existing lexicon, and are only secondarily used to form new words. It is important to realize that the 'organizing function' is not just a question of parsimony, but that it is important to explain morphological facts: word formation is based on generalizations in the existing lexicon.[15]

Under this view it is to be expected that morphological productivity is not a question of 'all or nothing'. Patterns can be more or less evident, depending on how many types of words they generalize over, how frequently those occur in the language, how noticeable the morphological effects are, and how uniform the semantic changes are that go along with them. It is important to realize that productivity is not just a question of the generality of the 'input' class. An affix can be restricted to attach only to a narrowly defined kind of stem of which there are not many instances, but still be completely productive within that domain.

To sum up, in order to deal with subregularities and exceptions in a principled way, it seems necessary to have a hierarchical model of some sort, because fine-grained distinctions about semantic and other word classes need to be made to capture the data. A traditional approach formulated in terms of word syntax or lexical rules is not sufficient. Both these approaches could be extended in the right direction by introducing hierarchies of affixes or lexical rules making the right fine-grained

[14]There is some independent evidence for the psychological reality of such patterns. Berko (1958) asked English speaking adults to form the past tense of nonsense words whose stems ended in the sequence *-ing*. 50% of them said **bang* or **bung* for the past tense of **bing*, and 75% made **gling* into **glang* or **glung*. This shows that adults do not necessarily use the most productive rule of their language and affix *-ed*, but instead may form the past tense in in analogy with existing irregular forms: most English verbs with stems ending in *-ing* have past tenses in *-ang* or *-ung*.

[15]This general conception of morphology is also advocated by Bybee (Bybee 1988 and Bybee 1994), who does not use a symbolic rule mechanism either. Salmons (1993) shows that such lexicon internal structure is relevant for at least one other aspect of synchronic grammar—gender assignment.

distinctions. But this would amount to a duplication of the hierarchy of words, which is independently needed to structure the lexicon, and which seems to be the primary level of classification, and probably the most natural starting point for children noting similarities among words.

For example, as we have seen, in the case of *ig*-adjectives there are many cases of lexicalized words where the suffix -*ig* is combined with nouns denoting potentially contaminating substances. These adjectives modify nouns denoting concrete physical objects, and they mean something like 'characterized by the presence of the contaminating substance (on the object)'. It is desirable to express the generalizations about this class of adjectives to structure the lexicon and in order to avoid having to repeat redundant information. Productive formation of *ig*-adjectives is crucially dependent on these existing word classes, and since the generalizations have already been expressed, it is easy to use them as schemata for productive word formation.

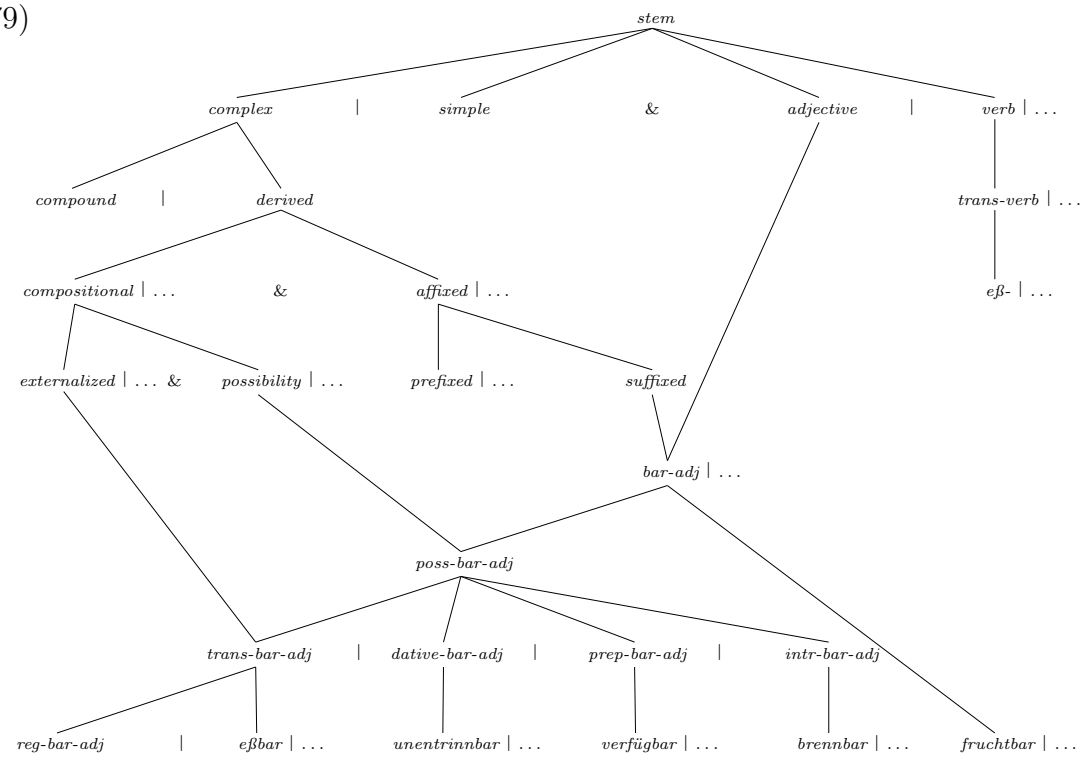
It would also be possible to express these fine-grained distinctions in a separate affix hierarchy. But this would result in an entry for -*ig* (among others) which says that it combines only with nouns denoting potentially contaminating substances and results in adjectives which are able to modify nouns denoting concrete physical objects, and which mean 'characterized by the presence of the contaminating substance (on the object)', *in addition to* the already existing word class of adjectives ending in *ig* able to modify nouns denoting concrete physical objects and meaning 'characterized by the presence of the contaminating substance (on the object)'. This is redundant, and it does not capture the fact that this affix entry exists and is productive precisely because the corresponding word class exists.

7.6 The Formal Approach

7.6.1 A Hierarchy of bar-Adjectives

(279) shows what a hierarchy could look like that addresses some of the problems mentioned above. It illustrates a simplified analysis of German *bar*-adjectives:

(279)



Explanation of the notation: '|' stands for disjunction (exclusive or), where either all disjuncts are specified, or it is indicated (by the dots) that there are (finitely many) more; '&' marks the boundaries between different dimensions of classification and always has wide scope over '|'. So, for example, what is said at the top of the hierarchy is that every stem must be either complex or simple, *and* either an adjective, or a verb, or belong to another of a finite number of categories which are not all listed here.

In this hierarchy stems are cross-classified according to various properties. It is assumed that stems in turn are parts of inflected words along the lines suggested in Kathol (1998), but the precise means of dealing with inflection is irrelevant for the treatment of derivation proposed here.

Since it is impossible to give a complete picture of such a hierarchy, some new notation has been introduced to mark where information is left out. This is especially important since I am assuming a closed-world interpretation, i.e. a given type does not have any more subtypes than the ones that are explicitly mentioned. Every linguistic object that is of a particular type has to be of one of the subtypes of that type, and

every object has to belong to a maximal type at the bottom of the hierarchy.

The specific stems at the bottom of the hierarchy are the lexicalized ones that need to be listed because of irregular properties or simply because they are conventionally known words (see Section 7.1). But they can be subsumed under the more general patterns one level above and only their idiosyncratic aspects need to be mentioned.

For the *bar*-adjectives from transitive verbs, there are so many lexicalized examples (which are, of course, not all shown here), and this pattern (*trans-bar-adj*) is so general and salient, that it is perceived as a productive rule. Because of the closed-world interpretation of the hierarchy, it is necessary to introduce a new type (*reg-bar-adj*) on the lowest level that inherits all the constraints from *trans-bar-adj*, and functions as the schema for productive word formation. This says in effect that such a *bar*-adjective is either one of the lexicalized ones, or any word that meets the constraints on the type *reg-bar-adj*, i.e. any word fitting this schema. Every appropriate stem of type *trans-verb* can be 'filled in' via unification, resulting in a new word with predictable properties.[16]

The approach can handle complex derivation because the stem itself can be complex, e.g. the verbal stem *werf* 'throw' can be prefixed by *ab-* 'off' to make the transitive verb stem *abwerf*, which is of the appropriate type to fit this schema, so the adjective *abwerfbar* 'throw-off-able' can be formed. This adjective is in turn of the appropriate type for *keit*-suffixation, so the noun *Abwerfbarkeit* can be formed. Recursive type constraints are formally explained in Carpenter (1992), Chapter 15.[17] Their morphological use here is parallel to their use for phrasal construction in HPSG—see e.g. Zajac (1992). One difference is that derivational schemata represent words, not phrases, and that many derived words are lexicalized whereas most phrases have to be constructed, although of course the derivational schemata can be used productively and there are lexicalized idiomatic phrases. A more significant difference is that *all* derivational affixes are introduced syncategorematically, whereas this is the exception

[16]So, formally a schema is just an ordinary complex type constraint, but in this chapter the term 'schema' is used to refer to those type constraints at the bottom of the hierarchy that can be used for productive word formation because they have an underspecified morphological base that can be 'filled in'.

[17]It would not be possible to implement this approach using standard macros or templates as e.g. in Shieber (1986), because these do not allow recursion.

for syntactic rules in most grammatical frameworks.[18]

Note also the introduction of a schematic subtype for *trans-bar-adj* provides a way of describing that this is a fully productive schema, although of course not explaining why—for that one would need a theory of productivity, and access to frequency information.[19] There is no schema for *bar*-adjectives from intransitive verbs on the lowest level. But the generalization *is* available, and one could imagine that various factors such as the degree of productivity of the class and the urgency of the need for a word could be the reason for making these subregular classes available as schemata for productive use.[20]

Let me explain the various functions of lexical type hierarchies. First of all, they structure the lexicon, by representing linguistically relevant subclasses of words explicitly. They thereby reduce redundancy in the representations for lexicalized (idiosyncratic or exceptional) words, and relate them to the rule, rather than just listing them. This seems desirable, because generalizations are expressed that would otherwise be lost, e.g. the relationship to passives and the semantic notion of possibility shared with other derived words. But it does not seem strictly necessary from a logical point of view, since there can only be a finite number of lexicalized words in a given language, which could in principle be listed. But simple listing would not only miss generalizations, e.g. the fact that all these words are adjectives and have similar semantics. It would also make it unnecessarily complicated to see these words as morphologically complex. It amounts to saying that the tiniest idiosyncrasy[21] in a

[18]The analogy to syntax also makes clear why it would be spurious to have an independent entry for the syncategorematically introduced material.

[19]It is hard to formalize what defines productivity. Semantic coherence, number of types and frequency of tokens, and predictability of properties all play a role. In some cases an affix might appear not to be fully productive only because very few lexical items match the constraints imposed on the stem, and those constraints may be hard to determine. These issues, in particular the role of frequency information, have been debated by Baayen (1992), van Marle (1992), Frauenfelder and Schreuder (1992), Clark (1993), and Bybee (1995). A formal approach to semi-productive lexical rules was recently developed by Briscoe and Copestake (1999).

[20]As a first approximation for increasing robustness in a computational system one could say that if a word cannot be found in the lexicon or analyzed by the productive schemata, all word-formation patterns with more than a certain number of subtypes can be tried. This would result in the system's ability to parse but not generate new formations like *verrottbar* 'rottable'.

[21]An example of this would be the word *einsetzbar* 'usable', which is transparently related to the transitive verb stem *einsetz* 'use'. The adjective nevertheless needs to be listed so it can be indicated

word has as a consequence that it is listed without any connection to its stem, that none of its properties are predictable, and that the existence of this lexical item does not affect which morphological rules there are in the language.

The patterns that arise by virtue of this structuring function can then be used to account for productive word formation: speakers use their knowledge of learned patterns to form new exemplars. It is not necessary to have a separate lexical rule mechanism for this purpose or use a word-syntactic approach.

And finally, these hierarchies can be used to describe word-formation that is analogical and not strictly rule governed. This is not yet fully worked out formally, but it is fairly straightforward to see what a processing regime might look like that uses existing (subregular) patterns to analyze unfamiliar words, and to produce new words when there is good motivation.[22] Ideally this would be based on a good theory of productivity, taking into account not only the number of lexicalized types and frequency information but also factors like the semantic coherence of the class.

7.6.2 The Productive Schema

The schema for the fully productive, regular *bar*-adjectives can be seen in (280).

that it means 'usable' while the verbal stem has many meanings, including the basic meaning 'put in' or 'insert', and more specialized meanings like 'apply' (force) and 'bet' (money), and 'risk' (one's life). Some of these senses could be made into *bar*-adjectives productively, but it is a fact about German that *einsetzbar* is normally used to mean 'usable'. In the Mannheim Corpus there are 38 occurrences of *einsetzbar*, all of which have this meaning—they are about the usability of medical substances, new technology, weapons, sources of energy, and personnel.

[22]For example, it is conceivable that people become more likely to form a slightly funny sounding *bar*-adjective if it is needed for coordination with other adjectives, or just because expressing the desired meaning with an adjectival semantics is important. E.g. when *breathable* bags have just been invented, or *renewable* sources of energy discovered, it is preferable to be able to express these new concepts concisely, especially if the syntactic alternatives are long and not very appealing. See also Clark and Clark (1979).

(280)
$$\left[\begin{array}{l}\textit{reg-bar-adj}\\ \textsc{phonology}\quad \boxed{1}\oplus \textit{bar}\\ \textsc{morph-b}\quad \left\langle\left[\begin{array}{l}\textit{trans-verb}\\ \textsc{phon}\quad \boxed{1}\\ \textsc{synsem}\,|\,\textsc{loc}\quad \left[\begin{array}{l}\textsc{cat}\,|\,\textsc{val}\,|\,\textsc{comps}\quad \langle \mathrm{NP}[\mathrm{acc}]_{\boxed{2}}{:}\boxed{5}\rangle\oplus\boxed{3}\\ \textsc{cont}\,|\,\textsc{lzt}\quad \left\langle \boxed{4}\left[\begin{array}{l}\textsc{act}\\ \textsc{und}\ \boxed{2}\end{array}\right]\right\rangle\end{array}\right]\end{array}\right]\right\rangle\\ \textsc{synsem}\,|\,\textsc{local}\quad \left[\begin{array}{l}\textsc{category}\quad \left[\begin{array}{ll}\textsc{head} & \textit{adj}\\ \textsc{valence} & \left[\begin{array}{ll}\textsc{subj} & \langle \mathrm{NP}{:}\boxed{5}\rangle\\ \textsc{comps} & \boxed{3}\end{array}\right]\end{array}\right]\\ \textsc{content}\,|\,\textsc{lzt}\quad \left\langle\left[\begin{array}{ll}\textit{possibility\_rel}\\ \textsc{arg}1 & \boxed{2}\\ \textsc{arg}2 & \boxed{4}\end{array}\right]\right\rangle\end{array}\right]\end{array}\right]$$

Through the various structure-sharings it specifies the relationships between the stem, which is the value of MORPH-B (morphological bases), and the whole complex word. For example, the semantics of the verb ($\boxed{4}$) is put into the scope of the possibility operator in the semantics of the adjective. $\boxed{2}$ is the UNDERGOER from the semantics of the verb (ACTOR and UNDERGOER are developed by Davis (1996) as part of his linking theory). This is to emphasize that the meaning of *bar*-adjectives is something like 'noun is capable of being verb-ed'—it is a notion of possibility qualified with respect to properties of the adjective's subject. Additional thematic restrictions on the UNDERGOER may be necessary, as discussed above. It is unclear how to write these down formally, but they are already part of the generalization over the lexicalized *bar*-adjectives from transitive verbs.

The schema also states that the CONTENT of the verb's accusative object ($\boxed{5}$) is the same as the CONTENT of the adjective's subject. And it specifies that further complements can be inherited, although it is debatable whether this should be dealt with syntactically (see Riehemann (1993)).

Phonologically, the affix is added to the phonology of the verbal stem. This does not always have to be simple concatenation as in the case of German *-bar*. The phonological constraints apply for each derivational step, automatically giving the effect of cyclic phonology.[23]

7.6.3 The Higher Level Generalizations

Not all of the information in the schema in (280) has to be given at this point in the hierarchy. This is because it makes sense to isolate the various parts of this constraint, to make it possible to express generalizations in the lexicon.

For example, it is true for all suffixed words that the suffix comes after the stem, as expressed in (281). The hierarchical lexicon enables us to express this kind of generalization. Therefore the argument in Selkirk (1982) against 'introducing affixes in the rules' as missing these generalizations does not apply.

(281)
$$\begin{bmatrix} \textit{suffixed} \\ \textsc{phonology} \ \boxed{1} \oplus \textit{suffix} \\ \textsc{morph-b} \quad \left\langle \begin{bmatrix} \textsc{phon} \ \boxed{1} \end{bmatrix} \right\rangle \end{bmatrix}$$

Note that the affix lives only in the phonology. It is not a sign and therefore not on the list of morphological bases. MORPH-B is a list because that is necessary for compounds.[24]

This appears to bring with it a locality effect: the information about previous affixes is not available to further affixations, or to phonology or syntax. But reference to the internal structure is needed in some kinds of word formation, as some affixes combine only with morphologically simple stems (*entflecken* ('de-stain'), **entfettflecken* ('de-grease-stain')) and others only with complex stems, sometimes even preferring particular affixes in their base (Hoeksema (1988), Aronoff (1976)). For

[23] See Orgun (1994) for data that cannot be handled by noncyclic approaches, and for an approach to phonology that seems compatible with the view of morphology expressed in this chapter.

[24] One can imagine that what happens when compounds develop into affixes is that the second element of the list 'drops out' of that list and only its phonology survives, since the word's semantic relationship to the complex word is no longer transparent.

example, -*ity* attaches most productively to adjectives ending in -*ic* -*al*, -*id*, and -*able* (*genericity*, *cardinality*, *hybridity*, *computability*). Where a limited amount of information about internal structure is needed (for example to make sure the -*ate* sequence that is truncated in English -*able* affixation is a formative (cf. **debable* vs. *operable*)),[25] it is available via the types. In most cases it is sufficient to know that a word has undergone affixation. But even more complex information, for example about the specific affixes required, can be made available in this way. Although affixes are not explicitly represented as part of the structure, the stems are, which makes the internal structure of complex words recoverable (in contrast to Anderson (1992)). But this information is not automatically available, which might explain why these phenomena are not common.

It seems independently necessary to allow words to be parts of other words for compounding, so there is nothing particularly strange about viewing the stem of a derived word as an independent word that forms part of the complex one. On the other hand, there is no need to view the affix as being a sign of its own.[26]

Semantically, *bar*-adjectives share the possibility aspect with other derived words ending in -*lich*.

(282)
$$\begin{bmatrix} \textit{possibility} \\ \textsc{synsem}\,|\,\textsc{local}\,|\,\textsc{content}\,|\,\textsc{lzt} \;<\; \textit{possibility\_rel} \;> \end{bmatrix}$$

This is not an uncommon situation: there are, for example, many affixes for agent nominalization, all of which have a related meaning (Beard 1990).

Syntactically, *bar*-adjectives have quite a lot in common with passives, in particular the fact that the object of the verb corresponds to the subject of the adjective (or passive):

[25] Carstairs-McCarthy (1992) has argued that the alternative phonological account offered by Anderson (1992) does not work for all cases.

[26] For an analysis of coordination involving affixes (e.g. *kostenlos oder -günstig*) consistent with this approach see Wiese (1996).

(283)
$$\begin{bmatrix} \textit{externalized} \\ \textsc{morph-b} \quad \left\langle \begin{bmatrix} \textit{trans-verb} \\ \textsc{synsem}\,|\,\textsc{loc}\,|\,\textsc{cat}\,|\,\textsc{val}\,|\,\textsc{comps} \quad \langle \text{NP[acc]:}\boxed{1}\ \ldots \rangle \end{bmatrix} \right\rangle \\ \textsc{synsem}\,|\,\textsc{local}\,|\,\textsc{category}\,|\,\textsc{valence}\,|\,\textsc{subj} \quad \langle \text{NP:}\boxed{1} \rangle \end{bmatrix}$$

The constraint on the subregular *prep-bar-adj* is that the semantics of the adjective's subject corresponds to the semantics of the verb's PP complement.

(284)
$$\begin{bmatrix} \textit{prep-bar-adj} \\ \textsc{morph-b} \quad \left\langle \begin{bmatrix} \textsc{synsem}\,|\,\textsc{loc}\,|\,\textsc{cat}\,|\,\textsc{val}\,|\,\textsc{comps} \quad \langle \text{PP:}\boxed{1}\ \ldots \rangle \end{bmatrix} \right\rangle \\ \textsc{synsem}\,|\,\textsc{local}\,|\,\textsc{category}\,|\,\textsc{valence}\,|\,\textsc{subj} \quad \langle \text{NP:}\boxed{1} \rangle \end{bmatrix}$$

Because of all the information available in the type *trans-bar-adj*, the lexical entries for lexicalized *bar*-adjectives can be minimal, and look like the one for *vergleichbar* in (285):

(285)
$$\begin{bmatrix} \textit{vergleichbar} \\ \textsc{morph-b} \quad \left\langle [\textit{vergleich}] \right\rangle \end{bmatrix}$$

Together with the inherited information, all the necessary phonological, syntactic, and semantic information about this complex word is available without stipulation:

(286)
$$\left[\begin{array}{l} \textit{vergleichbar} \\ \textsc{phonology}\quad \textit{vergleichbar} \\ \textsc{morph-b}\ \left\langle \left[\begin{array}{l} \textit{vergleich} \\ \textsc{phon}\quad \textit{vergleich} \\ \textsc{synsem}\,|\,\textsc{local}\ \left[\begin{array}{l} \textsc{cat}\,|\,\textsc{val}\,|\,\textsc{comps}\quad \langle \mathrm{NP}[\mathrm{acc}]_{\boxed{2}}{:}\boxed{5}\rangle \oplus \boxed{3}> \\ \textsc{cont}\,|\,\textsc{lzt}\ \left\langle \boxed{4}\left[\begin{array}{l} \textit{\_compare\_rel} \\ \textsc{act} \\ \textsc{und}\quad \boxed{2} \end{array}\right]\right\rangle \end{array}\right] \end{array}\right]\right\rangle \\ \textsc{synsem}\,|\,\textsc{local}\ \left[\begin{array}{l} \textsc{category}\ \left[\begin{array}{ll} \textsc{head} & \textit{adj} \\ \textsc{valence} & \left[\begin{array}{ll} \textsc{subj} & \langle \mathrm{NP}{:}\boxed{5}\,\rangle \\ \textsc{comps} & \boxed{3} \end{array}\right] \end{array}\right] \\ \textsc{content}\,|\,\textsc{lzt}\ \left\langle \left[\begin{array}{ll} \textit{possibility\_rel} & \\ \textsc{arg}1 & \boxed{2} \\ \textsc{arg}2 & \boxed{4} \end{array}\right]\right\rangle \end{array}\right] \end{array}\right]$$

Note that such an entry does not exist for non-lexicalized *bar*-adjectives such as *ziehbar*. But that word can be productively formed by unifying in the first part of the *reg-bar-adj* schema with the transitive verb stem *zieh*.

7.6.4 Zero-Derivation

In the constructional approach to morphology zero-derivation can be dealt with in a much nicer way than in word-syntactic systems. Because affixes do not have lexical entries, it is not necessary to assume lexical entries for suffixes without phonology. The only difference between a derived and a zero-derived form is that in the latter there is no effect of the derivation in the phonology.

7.6.5 Properties of the Approach

The proposed constructional approach expresses the relationship of subregularities and exceptions to the 'rule' and makes precise exactly which properties are shared in each case. Semiproductive patterns are available, along with information about the number of subtypes they have. This can be seen as the beginning of an account of degrees of productivity. Statistical information cannot be exploited in the standard HPSG formalism, but there is work on integrating probabilistic information with a typed feature structure system—see e.g. Brew (1995). So this kind of information could eventually be available for use in computational systems as well.

Affixes are not treated as signs, and they have no syntax and semantics; that is, they do not mediate syntactic and semantic properties between the 'input' and the 'output' category by first making these properties their own, as was the case in the approach in Krieger and Nerbonne (1993) in which the *-bar* suffix had all the syntactic and semantic properties of a *bar*-adjective. From this, it also follows that affixes are not heads. But the fact that most suffixes are category specific does not need to be expressed explicitly—it is hard to imagine a word formation schema which is general enough to match more than one major part of speech, yet specific enough to include a predictable syntax and semantics. Furthermore, all that is 'determined' by morphological heads is category (Zwicky 1985, Bauer 1990), which is easy to figure out even without any affixes (e.g. in zero-derivation). Also, affixes with a very specific semantics will be compatible only with one category—and the ones that are more flexible (e.g. diminutives) seem to be exceptions anyway.

The information embodied in the schemata is split up at one level into syntactic, semantic, and morphological information, and even further within that. This captures generalizations at higher levels which would otherwise be lost—some of the generalizations are useful for some other morphological phenomena.[27] For example,

[27]This should be constrained by a general theory of what a linguistically relevant generalization/class of words is. It is not quite as unrestricted as it may seem, since not all generalizations that are logically possible contribute to a simplification of the lexicon. Forming a class of words beginning in the letter 'a' for example would not reduce the amount of information that needs to be listed. But not all generalizations that reduce redundancy may be linguistically relevant, and further criteria need to be developed, taking into account significant clusterings of information, and whether a generalization is required in more than one place in the grammar.

the syntactic generalizations could be used for the passive construction and the semantic ones for *lich*-adjectives. In this case the morphological properties are not needed elsewhere, but for other affixes there are phonological effects that are independent of the semantics. Hence this approach expresses modular generalizations lost in rule approaches.

Inheritance is monotonic; that is, no information specified at a higher level can be overridden by more specific information lower in the hierarchy. This seems desirable from an acquisition point of view: if all information were defeasible it would be unclear how the schemata would be formed. Further potential problems with default systems have been discussed in Calder (1991) and Kilbury (1993). The price that has to be paid is a larger number of types in the hierarchy. It remains to be seen whether recent nonmonotonic logics that can distinguish between strict and defeasible information might be able to deal with the acquisition problem and avoid wrong classifications of new items to be entered into the hierarchy, by making clear what the defining characteristics are. An example of a potential problem for classification are the *bar*-adjectives from intransitive verbs—these have more features in common with the *lich*-adjectives than with the regular *bar*-adjectives. Meanwhile, it makes sense to pursue the monotonic approach further because it is actually more constrained, since none of the generalizations emerging from the data can be dropped or changed for the productive rule.[28]

Pinker (1999, 1997 and earlier work) argues that there is neurological evidence for a distinction between regular and irregular past-tense. There are patients who are able to process irregular past-tense forms but not regular ones, and there is also imaging data showing that different regions of the brain are active for these tasks.

If this turned out to be true for productively formed vs. irregular *derived* words as well,[29] that would not be inconsistent with the theory proposed here. Productively forming a word involves additional mechanisms not needed for the retrieval of

[28]Note that in the idiom approach defaults are used for quite a different purpose.

[29]This is hard to assess experimentally since many frequent words are lexicalized even though all their properties are predictable from their parts and the relevant schema, and it varies from speaker to speaker for which words this is the case. Note also that the data are not so clear even for inflection, because it is not obvious whether a whole verb or a verb in a particular tense should be considered the domain of irregularity.

lexicalized word forms, so whatever brain functions are used for locating the stem in memory and unifying it into the schema could be those that are impaired.

7.7 Hebrew Derivational Morphology

A templatic approach to morphology like the one described above for German and English is even more obviously needed to account for Hebrew derivational morphology. Semitic languages like Hebrew have a method of word-formation in which a fixed set of vowels is intercalated with root consonants. This cannot be captured in a word-syntactic approach. The effect could be achieved by a rule-based approach; however this would obscure the systematic and templatic nature of the relationship. A further reason why Hebrew provides interesting data in support of my approach is that sub-productive patterns abound (Clark and Berman 1984, Bolozky 1999).

7.7.1 Non-Concatenative Morphology

As an example of the non-concatenative nature of Hebrew word formation, consider the formation of instrument nouns using the *maCCeC* pattern. Such instrument nouns can be formed from any three-letter root by sticking these three root letters in the appropriate 'slots' in the pattern. In this case, the pattern consists of the sequence *ma* followed by the first and second root consonants followed by *e* followed by the third root consonant. As a lexicalized example, the instrument noun *masreq* ('comb') consists of the *maCCeC* pattern and the verbal root *s-r-q* ('comb'). As a productive example, the verbal root *s-d-r* ('order') can be made into the noun *masder* ('instrument for ordering').

The representation in (287) shows what such a derivational pattern of Hebrew would look like in my approach.

(287)
$$\begin{bmatrix} \textit{maCCeC-instrument} \\ \text{PHONOLOGY } \boxed{4} \\ \text{MORPH} \left\langle \begin{bmatrix} \text{ROOT} & \begin{bmatrix} \text{CONS} \left\langle \boxed{1}, \boxed{2}, \boxed{3} \right\rangle \\ \text{CONT } \boxed{5} \end{bmatrix} \\ \text{SKELETON } \boxed{4} \left\langle \text{m, a, } \boxed{1}, \boxed{2}, \text{ e, } \boxed{3} \right\rangle \end{bmatrix} \right\rangle \\ \text{SYNSEM} \mid \text{LOCAL} \mid \text{CONT} \mid \text{LZT} \left\langle \begin{bmatrix} \textit{instrument\_rel} \\ \text{PURPOSE } \boxed{5} \end{bmatrix} \right\rangle \end{bmatrix}$$

Similar approaches can be found for Hebrew adjectives of enablement in Melnik (1999), and for Sierra Miwok morphology in Bird and Klein (1994). In contrast to Bird & Klein, I am not structure-sharing the phonology of the root with that of the whole pattern, which I find counterintuitive. I am also assuming that the feature SKELETON is a morphological and not a phonological feature, at least for derivational morphology. So there is room for a theory of morphophonology to deal with phenomena like assimilation. Unlike Melnik (1999), I do not make the CONTENT of the pattern part of the morphology, instead the whole *maCCeC-instrument* pattern has that meaning.

7.7.2 Sub-Productive Patterns

Now let us consider an example of sub-productive patterns in Hebrew. There are a least 9 different ways of forming instrument nouns: with the suffixes -*an*, -*er*, -*iya*, and -*on*; and with the patterns *maCCeC*, *maCCeCa*, *meCaCeC*, *CoCeC*, and *CaCaC*. None of these is the 'preferred' method as can be seen in Table 7.1. Bolozky (1999:125-132), uses four different measures for productivity. By 'coinage' he means the forms 50 speakers produced in a production experiment, and 'judgment' are the forms the speakers picked from a list of options. 'Corpus' frequency is the frequency of each type of form among the hapaxes in a 112,000-word corpus, and 'dictionary' frequency is the frequency of these forms among words found in a newer dictionary but not an older version of the same dictionary.

	Coinage	Judgment	Corpus	Dictionary
-an	9.2%	14.0%	20.0%	7.6%
-er	2.8%	9.6%	0.0%	3.0%
-iya	23.6%	14.8%	13.3%	1.5%
-on	8.8%	4.8%	20.0%	4.6%
maCCeC	14.0%	11.6%	33.3%	39.4%
maCCeCa	3.6%	13.6%	6.7%	25.8%
meCaCeC	15.6%	13.6%	0.0%	4.6%
CoCeC	0.0%	0.0%	0.0%	3.0%
CaCaC	0.0%	0.0%	0.0%	10.6%

Table 7.1: Sub-Productive Derivational Patterns in Hebrew

While the *CaCaC* and *CoCeC* forms may not be productive for these speakers, the first seven forms are, with none of them dominating the derivation of instrument nouns.

Further evidence that native speakers know and use more than one form for instrument nominalization comes from Clark and Berman (1984). Table 7.2 gives the percentages of correctly identified verb roots for instrument nouns from a comprehension experiment with children and adults.

	Age 3	Age 4	Age 5	Age 7	Age 11	Adult
-an/CaCCan	58%	67%	86%	92%	92%	100%
maCCeC	29%	64%	48%	57%	71%	100%
maCCeCa	40%	57%	74%	81%	86%	100%

Table 7.2: Instrument Nominalization Comprehension Experiment

In a corresponding production experiment, the speakers also relied on compounding, suppletives (the use of existing words), and 'benoni' i.e. conversion from present tense verb forms, which include *CoCeC* and *meCaCeC*. The results from this experiment are summarized in Table 7.3.

In another experiment from Clark and Berman (1984), in which college-educated speakers were asked to write down their coinages for instrument nouns. 24% of the

	Age 3	Age 4	Age 5	Age 7	Age 11	Adult
-an/CaCCan	10%	43%	34%	36%	57%	53%
other patterns (*maCCeC, ...*)	1%	8%	3%	6%	5%	22%
benoni (*CoCeC, meCaCeC, ...*)	18%	5%	1%	10%	8%	15%
compounds	5%	8%	18%	30%	20%	2%
suppletives	33%	26%	38%	16%	9%	8%
don't know	33%	11%	7%	4%	1%	1%

Table 7.3: Instrument Nominalization Production Experiment

coinages were -*an* forms, and 30% were *ma-* forms.

In the constructional approach one would expect this 'messiness' given the variety of lexicalized instrument nouns in Hebrew. This is because in my approach, the productive 'rules' are seen not as separate from the existing lexicon, but as generalizations from it. Therefore, the prediction is that speakers use whichever patterns they have observed in the existing lexicon and in particular with recent innovations, with constraints being based on properties shared by these lexicalized words. So, because of the variety of forms that lexicalized instrument nouns take, it can be expected that speakers form different generalizations, depending on exactly which of these they know, how much they have analyzed them, and what kind of generalizations they have formed.

For example, it is possible that on the basis of having realized that they know many different kinds of kitchen tools of the form *maCCeC*, some speakers might choose to form novel nouns for kitchen tools in this pattern. Other speakers might instead choose the *maCCeC* form only with verbs from certain inflectional classes, because this is common among the words they know. Phonological similarity to other roots from which there already exist established instrument nouns may also play a role.

In addition, speakers' choices may be influenced by the semantic transparency of the word-formation device, i.e. they may prefer those that are specific to instrument nominalization, which again depends on knowledge of the patterns in the existing lexicon. Normative rules and speakers' intuitions on which word-formation devices

other speakers in the community actually use also play a role for speakers who are attuned to these things and choose to follow that lead.

The approach cannot predict exactly which form a given speaker will produce for a given root, because there are too many unknown variables about that speaker's mental lexicon. This is consistent with the nature of the data as described above. A rule-based approach, by contrast, would predict that there are a small number of well-delineated rules for making instrument nouns which bear no relationship to the forms of lexicalized instrument nouns.

7.8 Summary

In the constructional approach to morphology, generalizations from existing words are expressed as explicit schemata, which primarily reflect how the lexicon is organized, and only secondarily account for productive word formation.

This approach can do everything word syntax can do, and more—it can handle exceptions, and productive rules, but also existing partially regular words and new subregular cases. This avoids redundancy in lexical specifications by making explicit the partial relationships that exist to the productive rule. And it becomes possible to exploit existing subregular patterns for semiproductive word formation. In addition the proposed constructional approach is capable of capturing generalizations that are lost in other approaches, e.g. the syntactic properties shared with passives. Furthermore, it can deal with zero-derivation and other nonconcatenative morphology such as that found in Hebrew.

The approach is even more important for other affixes, such as the German *ig*-adjectives discussed in Section 7.2.3, which exhibit more irregularity and analogy. It is also needed for the semi-affixes discussed in Section 7.2.4 and it can express the generalizations at various levels of generality needed for compounding.

The data in Berko (1958) about past tense formation by analogy (see Section 7.3) suggest that an approach which is able to handle subregularities might also be needed for inflection, although this is not as crucial as for derivation. The proposed approach should generalize to inflection with some appropriate modifications.

It remains to be seen how much sense the approach makes for other languages. Some related work has been done already on Sierra Miwok morphophonology (Bird and Klein 1994), Korean inflection (Kim 1994), and Mohawk noun incorporation (Malouf 1994). Also, Bybee (1995) has shown that a lexicon-based view of morphology is appropriate for various languages, and is in fact needed to explain some of the data.

This constructional approach seems to be a natural way in which to make precise the old idea of 'redundancy rules', which goes back to at least Stanley (1967).[30] The idea behind redundancy rules is that their primary role is to give structure to the existing lexicon, and only secondarily to provide a means for word formation. Precisely this idea is captured by the proposed approach, which in addition can relate lexical rules to each other and account for subregularities.

The approach is a natural extension of the same formalism HPSG uses in its syntactic and lexical analyses in general, while it is not clear that a similar incorporation of the idea of 'redundancy rules' into other theories, such as the Minimalist Program, is possible.

[30] Jackendoff (1975) and Anderson (1992) have argued for this view of lexical rules. Bochner (1988) has given it an alternative formalization, which has recently been taken up again by Ackerman (1995).

Chapter 8

Analogy between Phrasal and Lexical Patterns

This chapter uses two lines of argument to show that the approach to idioms and the approach to derivational morphology presented above should be seen as instances of one general, lexicon-based, wholes-with-parts approach. The first part points out all the various analogous properties of the approaches, and the second part shows that there are several phenomena that are 'in between' and cannot be straightforwardly classified as either phrasal or lexical.

8.1 Shared Properties of the Approaches to Idioms and Morphology

In both cases, idioms and derived words, it was shown that there is a need for representing lexicalized instances. In the case of idioms, these are the canonical forms that were discussed in Chapter 3; in the case of derivational morphology, they are the high-frequency conventionalized derived words.

These lexicalized instances are wholes with parts in both cases: phrases with information about the words they contain, and words with information about the stems and affixes they contain. This captures the fact that the parts (idiomatic words

and affixes) are bound and cannot occur by themselves. The representations for the wholes are also necessary to provide a locus for the semantics of non-decomposable idioms and for semantically drifted meanings of complex words. In fact, these special meanings would not be expected to arise if there were no representation for these patterns in the grammar to which they could attach.

It was also argued that these lexicalized instances need to be classified and arranged in an inheritance hierarchy to avoid massive redundancy of predictable information. In the case of idioms this results in a very underspecified pattern which is consistent with all the possible variations of the idiom; in the case of morphology it is a very underspecified pattern which contains just the affix plus any properties of the stem shared by the lexicalized instances.

In both cases it is then argued that these underspecified patterns can account for the variation of idioms and for productive word formation, respectively. There is no need for an additional mechanism for these purposes. In both cases it is not possible to just recruit ordinary syntax for the job: idioms can involve more than the head-argument relationships typically expressed in lexical entries for verbs, and syntax does not come with a mechanism for restricting the occurrence of specific items which are subcategorized for, i.e. nothing prevents them from occurring by themselves. And 'word syntax' is not the same as syntax. Also, any additional mechanism that might be proposed is not only redundant but also unmotivated because it is not related to the existing lexicon, so there would be no explanation for its existence.

8.2 Between Lexical and Phrasal

In this section I look at phenomena 'in between' phrasal and lexical and show that a constructional approach is needed for them as well.

8.2.1 Compounding

There are two main points to be made about compounds in this context. The main point is that there are stem compounds, which are clearly a morphological phenomenon, and word compounds, which are of a more phrasal nature. Yet they also have many properties in common and should be given a similar analysis. This will be discussed in more detail below.

Another point, also made by Jackendoff (1997), is that compounds and idioms actually have several properties in common. Both compounds (*cranberry*) and idioms (*running amok*) can contain items that do not exist independently. Both compounds (*aloha shirt*) and idioms (*set foot*) can contain items in the wrong category. And both compounds (*sweetbread* (pancreas served as food)) and idioms (*kick the bucket*) can contain items that do not contribute to the meaning. So both compounding and idioms require a system to override some of the properties these items would otherwise be expected to have.

Most compounds are clearly lexical in the sense that a new stem is formed out of two other stems. This is quite clear in many languages, e.g. German. The clearest examples in German involve verbal stems, as in *Waschmaschine* 'washing machine', where *wasch* is a stem and cannot occur by itself without an inflectional ending. In English it is impossible to tell for most compounds in which the first part is singular whether or not they are stem compounds. The clearest evidence for stem compounds in English are examples like *pant leg*, *pajama party*, and *scissor cuts*, where the singular forms do not occur independently (**pant*, **pajama*, **scissor*).

However, some compounds have parts that look as if they are inflected, i.e. they are not stems but words. These can be found in many languages, including English (*sports page*), and German (*Bücherstube* 'books-room'). The English examples have been treated as compounds containing words, e.g. in the OT literature (Frank et al. 1998).

In a corpus study of the parsed Wall Street Journal corpus from Treebank II there are 1325 examples that are tagged as a plural noun (NNS) followed by a singular noun (NN), and 690 examples of a plural noun followed by another plural noun.

One interesting fact is that in many cases there is a good reason for the plural, i.e. either the noun only occurs in the plural (e.g. *electronics concern*, *savings account*)

or the singular noun means something else (e.g. *futures contract*). One might argue that the fact that these plural items need independent lexical entries explains why they show up in these compounds. This is true, however the items are still *words* and therefore necessitate a word-compound analysis.

There is also a good reason for the plural in cases where the singular noun would be ambiguous (e.g. *options strategist*, *sales unit*), or where it has a more frequent non-noun reading (e.g. *metals group*,[1] *rights law*). But in these cases there is no need for a lexical entry for the plural form because it can be derived from one of the meanings of the singular form. So independent lexical entries for plurals do not work as an explanation for these cases.

Furthermore, there seem to be some cases where none of the above applies, i.e. there does not seem to be a good reason for the plural in *appeals court*, *appropriations clause*, *capital-gains tax*, *sports car*, *creditors committee*, *mergers adviser*, *weapons maker*, *workers union*, *teethmarks*, or *systems analyst*.

While the plural often makes sense semantically, it is not used in all environments where this is the case. In fact, a plural is the exception even in these cases: **trials court*, **denials clause*, **luxuries car*, **professors committee*, **shoes maker*, **students union*, **computers concern*, **stocks account*, **orchids collection*, **films director*, etc. Recall that there are even examples like *pant leg* and *pajama party* where the singular has to be used (**pants leg*, **pajamas party*, **scissors cuts*) even though the nouns by themselves occur only in the plural.

At least in some of the cases, the motivation for the plural might be by analogy with similar things, like *earnings tax*, *sales tax* → *capital-gains tax*, or *arms control* → *weapons control*. Also, *workers union* might originally have been *workers' union*. However, Quirk et al. (1985:17.110) show that such genitive premodifiers are not ordinary possessives.[2]

Whether or not there is a good reason for the plural, they are clearly plurals, which necessitates a word-compound treatment. The strongest support for this comes

[1]This compound occurs in expressions like *the metals group of Barclays Bank Plc* and *the Swedish mining and metals group*.

[2]Consider the difference between *he joined the supportive (workers' union)* vs. *he joined his (supportive friends') union*.

from examples where the plural is irregular yet a singular exists and does not have a different meaning, e.g. *alumni association*, *bacteria recipients*, *women executives*, and *freshmen applicants*. In these cases a possessive reinterpretation is not possible.

There are also exocentric compounds in English like *court-martial* and *attorney general* whose plural can be realized on the first part of the compound: *courts-martial*, *attorneys general*.[3]

For German it has been argued that synchronically these are not inflected forms but that there is a special 'glue morpheme' involved. These glue morphemes certainly exist, but in some cases it is nevertheless desirable to indicate that the choice of 'glue morpheme' is not random but corresponds to the plural of that particular word. The most striking case is when irregular inflection is preserved even in productive compounding. For example, I found the words *Gäste-Account* ('guests account') and *Gästezugang* ('guests access') on a web page. These words must have been productively formed fairly recently. Nevertheless they use the irregular plural for 'guests', and it would not be possible to use any of the other glue morphemes (**Gasten-Account*, **Gasts-Account*, **Gaster-Account*), although it is possible to form the compound *Gast-Account* without any glue morpheme. The best way to capture this is by reference to the irregular plural form of the word *Gast*.

In other languages the case for inflected words as parts of compounds is even clearer. For example, Stump (1991) shows that inflection of Breton compounds can be reflected on their internal parts. This is like the English plural *courts-martial* except that there are examples with irregular plurals as well.

In Hebrew the data are even more interesting: both plural and definiteness markers are always inserted in the middle of compounds. The plural is marked on the first part of the compound, which is the head. And the definiteness marker is a prefix to

[3]One might think that these plurals are only used in very formal registers, but a Google search on 5/27/2001 found that these plurals are much more frequently used on the web than their counterparts with the plural realized on the second part. The form *courts martial* is used 9 times more frequently than *court martials* (22900 vs. 2530 hits) even though some of the latter may be verbs, and *attorneys general* is used 6 times more frequently than *attorney generals* (97900 vs. 15600 hits). The difference is not so striking in the even more colloquial language in the Deja/Google newsgroup archive, but the forms *courts martial* and *attorneys general* are still more frequent (2.5 times and 1.3 times, respectively).

the second part of the compound.

(288) *mechonit masa*
car load
'(a) truck'

(289) *mechonit hamasa*
car DEF-load
'the truck'

(290) *mechoniyot masa*
car-PL load
'trucks'

(291) *mechoniyot hamasa*
car-PL DEF-load
'the trucks'

This is a regular pattern, i.e. the Hebrew equivalent of *the grocery stores* is *chanuyot hamakolet* ('stores the-grocery'), *the aerograms* is *igrot haavir* ('letters the-air), *the breakfasts* is *aruchot haboker* ('meals the-morning') and so on. For most feminine singular and masculine plural nouns there is a special ending when they are the first part of a compound, but for feminine plural nouns the exact same form is used, so they are clearly words and not stems. Another reason they need to be words is that otherwise the syntax would not be able to insert the definiteness marker.

So it is clear that there exist compounds which are made up of words, and other compounds that are made up of stems. However, many of the other properties of these compounds are similar, e.g. the types of semantic relationships between the parts and the ways in which they are lexicalized. For example, the semantic relationship between *savings* and *account* in *savings account* is the same as that in *cash account*, i.e. 'account where one keeps one's savings/cash'. And the relationship in *sales tax* is the same as in *income tax*, i.e. 'tax one pays on sales/income'. And an *alumni association* is like a *student association* in that it is an 'association of alumni/students'.

Also, as noted above, parts of lexicalized word-compounds usually have only one meaning even when the underlying words have multiple senses. For example, *sales* in

sales unit refers to things being sold, not being on sale at a discount. And *options* in *options strategist* refers to stock options, not choices in general. This is something that is true for stem compounds as well: a *bank account* is an arrangement with a financial institution, not a report about a slope next to a river, and a *bear market* is not a place where bears are sold. For both word-compounds and stem-compounds these other meanings are available to be formed productively, but there may be blocking effects in both cases.

This shows that phenomena on both sides of the boundary between lexical and phrasal items require similar treatment. When morphological markers are absent it is not even clear where the boundary is.

(292) $\begin{bmatrix} \textit{stem\_compound} \\ \textsc{morph-b} \quad \left\langle [\textit{stem}] \ , \ [\textit{stem}] \right\rangle \end{bmatrix}$

(293) $\begin{bmatrix} \textit{word\_compound} \\ \textsc{morph-b} \quad \left\langle [\textit{word}] \ , \ [\textit{word}] \right\rangle \end{bmatrix}$

It is not clear whether the type *word_compound* should be a subtype of *word* or *phrase*, given that it has some properties of both. The external distribution of a *word_compound* is like that of a word, while internally it consists of other words, like phrases do (see Malouf 1998).

Some word-compounds like *ombudsman* or *passers-by* contain words that do not exist independently with the same meaning. To deal with this one can use some of the same mechanisms that were developed for idioms in Chapter 5. In particular, a mechanism is needed to make sure that the phonological and inflectional properties are related as necessary, and it is necessary to capture the fact that these words cannot occur with their special meaning outside of the compound.

In fact, it is possible to use the approach developed for idioms, given that *word_compound*s contain words. For example, a compound like *passers-by* would look like this:

(294)
$$\left[\begin{array}{l} \textit{passers\_by} \ \& \ \textit{word\_compound} \\ \textsc{morph-b} \ \left\langle \boxed{1} , \boxed{2} \right\rangle \\ \textsc{words} \ \left\{ \begin{array}{l} \boxed{1} \ [\textit{passers}], \\ \boxed{2} \left[\ldots\textsc{lzt} <\textit{empty\_rel}>\right] \overset{\leq}{\sqcap} \left[\ldots\textsc{lzt} <\textit{by\_rel}>\right] \end{array} \right\} \\ \ldots \ \textsc{lzt} <\textit{\_passers\_by\_rel}> \end{array}\right]$$

8.2.2 Separable Prefix Verbs in German

Separable prefix verbs in German look like a derivational morphology phenomenon in some syntactic contexts. For example in the infinitive *ankleben* ('glue on'), *an-* ('on') is a prefix and *kleb* ('glue') is a stem (*-en* is the infinitive ending). These are called synthetic occurrences.

However, when an inflected form of the verb occurs in a main clause that does not contain an auxiliary, the 'prefix' actually separates from the stem and occurs by itself at the end of the clause. These are called analytic occurrences.

(295) *Er muss das Plakat ankleben.*
he must the poster on-glue
'he must glue the poster on'

(296) *Er klebt das Plakat heute an.*
he glues the poster today on
'he glues the poster on today'

It has been clearly established in the literature (e.g. Fleischer and Barz (1992), Ackerman and Webelhuth (1998)) that the synthetic occurrences are really one morphological word. They exhibit the intonation of compounds and no other words can ever intervene between their parts.

It is also established (Fleischer and Barz 1992:29-30) that it is not possible to treat the separated prefixes as instances of the prepositions they are often homonymous with. The reasons for this are that they often do not have the same meaning as these prepositions, and show various affix-like properties such as occurring in clusters of

related words. For example, the preposition *an* means 'on' or 'at', and *an-* can have those meanings as a separable prefix. But it can also have various other meanings. For instance, there is a whole set of verbs in which the *an-* signals that the action is started but not completed: *braten* means 'fry' and *anbraten* means 'briefly start to fry', *schneiden* means 'cut' and *anschneiden* means 'start to cut', *zahlen* means 'pay' and *anzahlen* means 'start to pay, make a down payment', and so on.

Furthermore not all of these prefixes correspond to a homonymous independent word. For example in the infinitive *einschalten* ('switch on'), *ein-* is a prefix that does not correspond to an independent preposition or adverb. Yet it separates in exactly the same way:

(297) *Er muss das Licht einschalten.*
he must the light SP-switch
'he must switch the light on'

(298) *Er schaltet das Licht ein.*
he switches the light SP
'he switches the light on'

Note that separable prefix verbs show the same irregular inflections as the verbs on which they are based, even when they are not semantically related in any way.

(299) *Der Mann wird fallen.*
the man will fall

(300) *Das Buch wird mir auffallen.*
the book will me SP-fall
'the book will attract my attention'

(301) *Der Mann fiel.*
the man fell

(302) *Das Buch fiel mir auf.*
the book fell me SP
'the book attracted my attention'

Furthermore, for inflected forms that normally have prefixes (e.g. *ge-* for past participles and the *zu-* for infinitives), these remain prefixes to the verbal part, i.e. result in what looks like infixes in the derived verbs:

(303) *Er muss das Licht einschalten.*
he must the light SP-switch
'he must switch the light on'

(304) *Er hat das Licht eingeschaltet.*
he has the light SP-INF-switched
'he switched the light on'

When a verb exceptionally does not take the *ge-* prefix in the past participle, the separable prefix verbs derived from it do not, either:

(305) *Er muss das Bett beziehen.*
he must the bed cover
'he must cover the bed'

(306) *Er muss das Kind einbeziehen.*
he must the child SP-cover
'he must involve the child'

(307) *Er hat das Bett bezogen.*
he has the bed covered
'he has covered the bed'

(308) *Er hat das Kind einbezogen.*
he has the child SP-cover
'he has involved the child'

There are two main issues that arise from these data. The first is how to represent the separated versions of the verbs in a satisfactory way, capturing the fact that the separated prefixes are separate units syntactically but not semantically. The second is how to capture the relationship between the separated and non-separated instances of these words. As Ackerman and Webelhuth (1998:322-324) have convincingly argued,

it is not satisfactory to treat them as separate lexical entries, even ones related by lexical rule. The lexicon would contain hundreds of predicates which all happen to have the same kind of defective paradigm, e.g. not being able to occur in subordinate clauses, and there would happen to be a lexical rule that applies to exactly this set of entries and which produces virtually identical entries which fill exactly the missing paradigm slots. This fails to express that the two lexical entries together function in the same way as a regular non-particle predicate in German. And there would be nothing to explain why there is not a single instance where the two entries have drifted apart in their meaning, which would be expected if there were separate lexical representations.

The first issue of how to represent the separated occurrences has an easy resolution in my approach, which is essentially like the treatment of idioms:

(309)
$$\left[\begin{array}{l} \textit{schalt\_ein\_analytic} \\ \text{WORDS} \left\{ \left[\begin{array}{l} \ldots\text{LZT} \ <\textit{empty\_rel}> \\ \ldots\text{COMPS} \ \ <\text{NP}> \oplus \boxed{1} \\ \ldots\text{TOPO} \ \ \textit{cf} \end{array}\right] \overset{\leq}{\sqcap} \left[\begin{array}{l} \textit{verb} \\ \ldots\text{LZT} \ <\textit{\_schalt\_rel}> \\ \ldots\text{VFORM} \ \ \textit{fin} \end{array}\right], \right. \\ \qquad\qquad \left. \left[\begin{array}{l} \textit{ein\_sep\_pref} \\ \text{PHON} \ <\text{ein}> \\ \ldots\text{SYNSEM} \ \ \boxed{1}\left[\ldots\text{LZT} \ <\textit{empty\_rel}>\right] \\ \ldots\text{TOPO} \ \ \textit{vc} \end{array}\right], \ldots \right\} \\ \text{C-CONT} \ <\textit{\_switch\_on\_rel}> \end{array}\right]$$

Here, the verb with the meaning *_schalt_rel* is a finite inflected form of the verb *schalten* except without its usual meaning, and *ein_sep_pref* is an item that has no meaning and does not have an independent lexical entry but occurs only with this particular type of separable prefix verb.[4]

Since I am not concerned with German word order in this dissertation I use the feature TOPO from Kathol (1995) to describe the ordering constraints. TOPO *cf* corresponds to the 'linke Satzklammer' of traditional German grammar, i.e. the V2

[4]Note that approaches requiring a separate entry for the separable prefix have to limit its occurrence to this context in some other way.

position in ordinary main clauses. TOPO *vc* corresponds to the 'rechte Satzklammer', i.e. the verbal cluster towards the end of a sentence (only followed by 'extraposed' items).

The second issue of how to relate the synthetic and analytic incarnations is trickier. Because of the fact that irregular inflections are preserved, it is necessary to relate the derived forms not to the verbal stem, but to the verb itself, with its potentially irregular inflections. The resulting representation for the synthetic forms of these words is:

(310)
$$\begin{bmatrix} \textit{schalt\_ein\_synthetic} \\ \text{PHON} \quad \boxed{1} \oplus \boxed{2} \\ \text{MORPH-B} \quad < \boxed{3} , \boxed{4} > \\ \text{WORDS} \quad \left\{ \begin{array}{l} \boxed{3} \begin{bmatrix} \textit{ein\_sep\_pref} \\ \text{PHON} \quad \boxed{1} <\text{ein}> \\ \ldots\text{LZT} \quad <\textit{empty\_rel}> \end{bmatrix}, \\ \boxed{4} \begin{bmatrix} \text{PHON} \quad \boxed{2} \\ \ldots\text{LZT} \quad <\textit{empty\_rel}> \\ \ldots\text{COMPS} \quad <\text{NP}> \end{bmatrix} \overset{\leq}{\sqcap} \begin{bmatrix} \textit{verb} \\ \ldots\text{LZT} \quad <\textit{\_schalt\_rel}> \end{bmatrix} \end{array} \right\} \\ \text{C-CONT} \quad <\textit{\_switch\_on\_rel}> \\ \text{TOPO} \quad \textit{vc} \end{bmatrix}$$

Now it is clear how the synthetic and analytic forms of these verbs are related: they have a common supertype specifying all the shared information, i.e. the fact that they are both phrases containing the same two words and having the same meaning.

(311)
$$\begin{bmatrix} \textit{schalt\_ein\_spv} \\ \text{WORDS} \quad \left\{ \begin{array}{l} \begin{bmatrix} \ldots\text{LZT} \quad <\textit{empty\_rel}> \\ \ldots\text{COMPS} \quad <\text{NP}> \end{bmatrix} \overset{\leq}{\sqcap} \begin{bmatrix} \textit{verb} \\ \ldots\text{LZT} \quad <\textit{\_schalt\_rel}> \end{bmatrix}, \\ \begin{bmatrix} \textit{ein\_sep\_pref} \\ \ldots\text{LZT} \quad <\textit{empty\_rel}> \end{bmatrix}, \ldots \end{array} \right\} \\ \text{C-CONT} \quad <\textit{\_switch\_on\_rel}> \end{bmatrix}$$

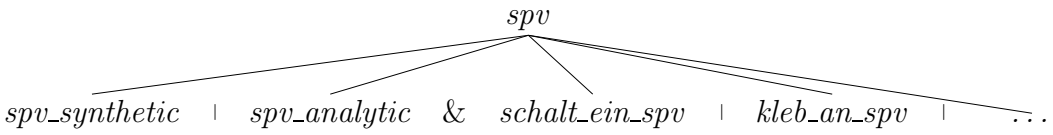
The synthetic subtype in (310) further specifies that the phrase is actually a phrasal compound with no other daughters, as the complete set of WORDS is given. The analytic compound in (309) does not specify how many other daughters this phrase contains, and instead specifies the syntactic relationship between the verb and the particle.

More generally, all separable prefix verbs have just one lexical entry meeting the constraints on the type *spv*:

(312)
$$\begin{bmatrix} \textit{spv} \\ \text{WORDS} \quad \left\{ \begin{bmatrix} \ldots\text{LZT} \ <\textit{empty\_rel}> \end{bmatrix} \overset{\leq}{\sqcap} \begin{bmatrix} \textit{verb} \end{bmatrix}, \right. \\ \left. \begin{bmatrix} \textit{sep\_pref} \\ \ldots\text{LZT} \ <\textit{empty\_rel}> \end{bmatrix}, \ \ldots \right\} \end{bmatrix}$$

This type has two partitioning subtypes, *spv_synthetic* and *spv_analytic*, which are cross-classified with the lexical entries, as in (313).

(313)



The constraints on these types are as in (314) and (315):

(314)
$$\begin{bmatrix} \textit{spv\_synthetic} \\ \text{PHON} \quad \boxed{1} \oplus \boxed{2} \\ \text{MORPH-B} \quad < \boxed{3} , \boxed{4} > \\ \text{WORDS} \quad \left\{ \boxed{3} \begin{bmatrix} \textit{sep\_pref} \\ \text{PHON} \quad \boxed{1} \end{bmatrix}, \right. \\ \left. \boxed{4} \begin{bmatrix} \text{PHON} \quad \boxed{2} \end{bmatrix} \overset{\leq}{\sqcap} \begin{bmatrix} \textit{verb} \end{bmatrix} \right\} \\ \text{TOPO} \quad \textit{vc} \end{bmatrix}$$

(315)
$$\left[\begin{array}{l} \textit{spv\_analytic} \\ \text{WORDS} \left\{ \begin{array}{l} \left[\begin{array}{l} \ldots\text{COMPS} \oplus \boxed{1} \\ \ldots\text{TOPO} \;\; \textit{cf} \end{array}\right] \sqsubseteq \left[\begin{array}{ll} \textit{verb} & \\ \ldots\text{VFORM} & \textit{fin} \end{array}\right], \\ \left[\begin{array}{l} \textit{sep\_pref} \\ \ldots\text{SYNSEM} \;\; \boxed{1} \\ \ldots\text{TOPO} \;\; \textit{vc} \end{array}\right], \ldots \end{array} \right\} \end{array}\right]$$

This analysis truly captures what the two incarnations of these separable prefix verbs have in common, while an analysis using disjunctions, like that in Ackerman and Webelhuth (1998), shown in (316), does not.

(316)
$$\left[\begin{array}{l} \text{SYNSEM} \left[\begin{array}{l} \text{CSTRUC} \left[\text{CR} \left[\begin{array}{l} \textit{vword} \\ \text{MS} \;[\; \text{LME} \;<\textit{rufen}> \vee <\textit{an},\textit{rufen}> \;]\end{array}\right]\right] \\ \text{VAL} \left[\text{PART} \left\langle \left[\text{CSTRUC} \,|\, \text{HD} \,|\, \text{MS} \;[\text{LME} \;<\textit{an}>]\right] \right\rangle \vee < \;\; >\right] \end{array}\right] \end{array}\right]$$

Because one can put a disjunction sign between totally different kinds of information, disjunctions do not necessarily express shared properties. In this example, it would be possible to add a synonym to the disjunction, in which case the relationship between the two incarnations of the separable prefix verb would be no more closely related to each other than they are related to the synonym. Or it would be possible to replace one part of the disjunction with a synonym, e.g. $<$*an*, *telefonieren*$>$. The fact that *rufen* and *an* appear twice in (316) is a coincidence and not required formally.

8.3 Summary

In this chapter I have pointed out all the properties that the constructional approach to idioms and the constructional approach to word formation have in common. I also have studied two sets of data that are 'in between' phrasal and lexical: compounding and separable prefix verbs in German. It was shown that a constructional approach is desirable for these data as well.

Chapter 9

Conclusions

In this dissertation I have argued for a general program of using a constructional approach for idioms and derivational morphology. I have also shown that the approach can be used for other aspects of grammar, such as syntactic constructions, and explored phenomena that straddle the lexical/phrasal boundary, such as compounds and separable prefix verbs in German. I have presented this approach in enough formal detail to indicate how it could be implemented in a computational HPSG system, although some discrepancies between the formalization and the intuitions underlying the approach remain to be resolved.

The approach views all these as complex patterns with sub-parts, as opposed to separate pieces and ways for assembling them. For idioms this means focusing on the phrasal level and viewing idiomatic words as parts of idiomatic phrases; for morphology, this means looking at the word level and viewing stems and affixes as parts of complex words. Neither the idiomatic words nor the affixes have an existence outside of these complex patterns, and their meaning is instead associated with the larger unit.

This approach is motivated by the fact that these larger patterns are based on the existing lexicon or 'constructicon'. Canonical forms of idioms and lexicalized instances of productive morphological patterns need to be represented in order to account for all of the systematicity of the data, although they are not needed if grammar is seen as merely describing the set of all possible structures without regard to the frequency

and manner of their use, including subtle pragmatic nuances. In both cases, idioms and morphology, there is also independent evidence for the need for these complex patterns.

In the case of idioms, the main motivation is that words cannot occur in their idiomatic meanings outside of the idiom, and that idioms can involve more than just combinations of head and arguments as usually seen in subcategorization. Also, in some semantically non-decomposable idioms such as *kick the bucket*, the phrasal pattern is needed to carry the semantics, because the meaning of the idiom cannot be broken down into parts associated with the individual words making up these idioms. Further evidence comes from the interaction of idioms and constructions.

In the case of derivational morphology, the main motivation for this approach is that the complex patterns are needed independently to structure the large inventory of lexicalized words, many of which are not fully compositional. A separate lexical rule mechanism would be redundant because the same patterns can be used for productive word formation. The patterns are also needed to explain subregular productivity, and to handle non-concatenative morphology. Further motivation comes from the fact that affixes are items which can never occur by themselves.

So a constructional approach is needed for both idioms and derivational morphology. The two have much more in common than being 'constructional', i.e. sharing a wholes-with-parts perspective. In both cases, part of the motivation for the larger construction is that the bound parts (affixes and stems, and idiomatic words) cannot occur by themselves. In both cases, the larger pattern is needed for conventionalized instances (derived words, and canonical forms of idioms). In both cases, the way the parts combine is not quite like anything else in the grammar: 'word syntax' is rather different from sentential syntax, and idioms involve more than the head-argument relationships usually handled via subcategorization. In both cases the larger pattern is also needed as the locus for special meanings—to describe the semantic drift of derived words, and the semantics of non-decomposable idioms. In fact, these special meanings would not be expected to arise if there were no representation for these patterns in the grammar to which they could attach. In both cases, productive formations (productively formed words, and varied idioms) need to be related to the

conventionalized ones. This follows immediately if the productive pattern is a generalization of the conventionalized ones. And in both cases, an additional unrelated mechanism is redundant.

The data and analysis in this dissertation can be seen as motivation for 'Experience-Based HPSG', in which representations are reinforced when used, complex items can be stored when they are interesting or unusual for any reason, and language processing consists largely of the retrieval and unification of these partially pre-assembled chunks. In this view grammar is the result of generalizing from the stored pieces.

9.1 Further Work - Experience-Based HPSG

In this dissertation I have argued that there is a need to include representations of the canonical form of idioms. One might claim that this is not necessary, because it is possible to generate these strings from the more underspecified representation that is needed independently to account for these idioms' variation. However, the evidence studied in this dissertation, specifically the convergence between the corpus data and the idioms' citation forms, suggests that information about canonical forms is part of knowledge of language. So a psycholinguistically plausible model must encode this information somewhere. HPSG provides a rich enough representational medium to allow this information to be encoded in the grammar, and in fact it has quite a natural place in the grammar as it refers to many linguistic properties (see also Jackendoff (1997)). In contrast, current theories of performance compatible with HPSG are too rudimentary to provide a satisfactory way of encoding information about canonical forms.

The reasons I gave for representing these canonical forms were based on corpus research and psycholinguistic evidence. My corpus study showed that about 75% of idiom occurrences are in their canonical form. This was compared to a baseline of non-idioms with similar meanings (e.g. *reveal* and *secret*), for which the most frequently found form (*reveal secrets*) only accounts for about 7% of the data and thus cannot be called "canonical form". On average the most frequent form of non-idioms accounts for about 16% of their occurrences, showing that the high frequency of canonical

form occurrences of idioms cannot be fully explained by semantic and other factors. This suggests that knowledge about the existence of these canonical forms must be represented as part of the knowledge of language.

Representations for canonical forms of idioms are to be expected in a usage-based approach (Langacker (1987), Langacker (1990), Barlow and Kemmer (2000)):

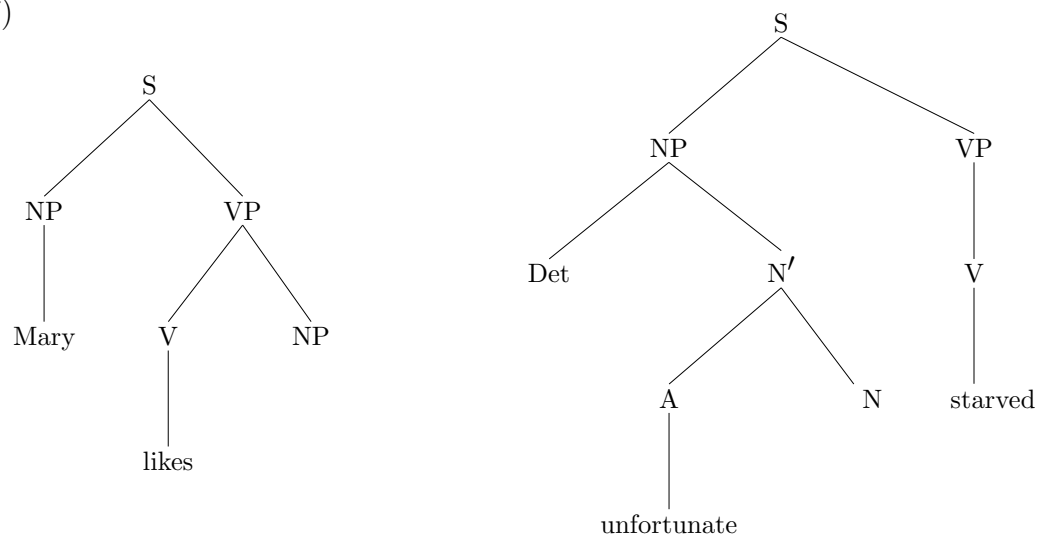
> The grammar lists the full set of particular statements representing a speaker's grasp of linguistic convention, including those subsumed by general statements. ... Speakers do not necessarily forget the forms they already know once the rule is extracted, nor does the rule preclude their learning additional forms as established units. (Langacker 1987:46)

Further evidence for this comes from psycholinguistic studies (e.g. McGlone et al. 1994) which have shown that idioms in canonical form are understood faster than their variants. It is generally accepted that retrieval from memory is faster in humans than other forms of processing. Therefore this psycholinguistic finding should be predicted by an approach which has representations for canonical forms of idioms which are quite specific and therefore require less processing because they can be retrieved from memory as a whole. HPSG argues for what Sag and Wasow (1999) call 'performance-plausible competence grammar', i.e. models of linguistic competence that can be embedded in models of language processing. On such a model, psycholinguistic results can influence the design of grammar and specific linguistic analyses (see also Bender (2001), Chapter 6).

The psycholinguistic evidence is consistent with HPSG viewed as a constraint-based system using unification to put together representations—the bigger the pieces, the less work unification has to do. However, this is not necessarily the case in standard HPSG processing architectures, because the parsers which are usually used are not equipped to deal with these phrasal representations. There are some HPSG systems like TFS (Typed Feature Structure System) which can handle the phrasal representations because they work by using type inference and general constraint-resolution mechanisms. In these systems it is possible that larger and more complete phrasal representations could lead to faster processing. But it is not guaranteed, as nothing ensures that these choices are explored first.

In this section I would like to discuss augmenting a system like TFS with heuristics like some of those used by Neumann (1997) and Bod (1998). In both of these approaches chunks of previous parses are stored. They are then used for various purposes, such as to speed up processing for subsequent parses, to help disambiguate, and to pick good candidates for generation. These approaches are not psycholinguistically plausible either, because they are not selective in which pieces of observed utterances they store. E.g. Bod (1998) unselectively stores trees with just some terminal nodes instantiated, such as the trees in (317).

(317)



These trees may or may not contain interesting semantic relationships, but it is implausible to assume that human listeners store large numbers of random pieces for every utterance they hear, and search though all of these stored pieces when trying to generate an utterance. Instead it is more likely that people preferentially store high-frequency patterns and complete constituents or other conceptual units. However, the mechanisms these approaches have developed for using the existing chunks to speed up processing might help make HPSG a better model of human sentence processing. For example, in Neumann's approach the phrasal templates are added as a passive edge to the chart parser/generator's agenda, and passive edges with a larger span are given higher priority.

The way I conceive of the mapping between HPSG and human sentence processing is as follows. I assume that HPSG representations, such as those for canonical forms of idioms, are a simplified model for whatever representations humans have for these forms. But how do people come to have these representations, and how are they subsequently used?

The first step is that all the representations that were used for each successful parse are reinforced: the representations of the lexical items as well as the general syntactic constructions become stronger. When an item is encountered that does not have the meaning which would be expected from the existing representations, or carries additional semantic or pragmatic information, that item is stored as well. Over time this will result in a multiply reinforced representation for the canonical form of idioms, and the formation of a generalization over any variations which may have been encountered.[1] The infrequent variations themselves will "fade away" eventually because of lack of reinforcement.

For collocations the picture is similar, except that it is not as easy to see why they are stored in the first place because they can be interpreted compositionally. By my definition collocations are fixed expressions made up out of two or more words which do have one of the meanings they can have independently, but which are conventionalized or established in this combination, e.g. *bear the brunt of*. Since their meaning is fully compositional there is no need to give them a special representation in order to make sure this string is part of the set of grammatical English expressions. But this would fail to predict that they are more frequent and more likely contain the words in the same senses than would be expected by chance. For example *to hold one's turf/ground* is not likely to mean 'grasp a piece of soil'. This is the sort of frequency fact that should be captured by grammar, as Bender (2001) argued convincingly.

Apart from the fact that the parts of collocations co-occur more frequently than would be expected there are other reasons why collocations are not entirely predictable from the point of view of the rest of the grammar. They often contain words that are not frequently used (like e.g. *bear the* ***brunt*** *of*), or are not frequently used in the particular sense they have in the collocation (like e.g. *more than NP* ***bargained***

[1] How much data is needed to form a generalization may vary from speaker to speaker.

for), and/or there is some additional pragmatic import associated with the use of the collocation (like e.g. in *past NP's prime*).

The second step is easy to see for human sentence processing: it is more efficient to retrieve (partially) pre-computed phrasal representations than to assemble them. In order to model that, one may have to use some of the methods and heuristics of Neumann (1997) and Bod (1998) to make sure the existence of these phrasal representations actually lead to faster processing, so that canonical forms are preferred.

Appendix A

List of the Corpora Used

1. **North American News Text Corpus** from the **LDC** (350 million words):
   New York Times News Syndicate, 7/94-12/96 (173 million words)
   Wall Street Journal, 7/94-12/96 (40 million words)
   Los Angeles Times & Washington Post, 5/94-8/97 (52 million words)
   Reuters News Service General & Financial News, 4/94 - 12/96 (85 million words)
2. **Wall Street Journal** from **Treebank II** (1 million words)
3. **Google Groups** newsgroup archive, more than 6 years of posts
4. **LEXIS-NEXIS General News**, more than 20 years of articles from Major Newspapers
5. **Frankfurter Rundschau, Donau Kurier, and VDI Nachrichten** from the **ECI/DCI Multilingual Corpus I** (50 million words)
6. **Mannheim Newspaper Corpus**, 1985-1988; 10,766,759 words
   17,297 *bar*-adjective tokens, 836 types
7. **Rosengren Newspaper Corpus** (Rosengren 1977), 1966/67; 2,976,894 words
   4346 *bar*-adjective tokens, 343 types (no context)
8. **Gelhaus Mannheim Corpus** (Gelhaus 1977)
   1206 *bar*-adjective tokens, 313 types (no context)

9. **Wendekorpus** (mostly Newspapers), 1989/90; 488,563 words
   661 *bar*-adjective tokens, 152 types
10. **Journal 'Germanistik'** 1991; 385,531 words
    343 *bar*-adjective tokens, 119 types
11. **Electronic Newsgroups**, ELWIS project, University of Tübingen, 1992/93; roughly 20 million words

Appendix B

List of the Dictionaries Used

1. NTC's American Idioms Dictionary, Second Edition, Richard A. Spears, Ph.D., NTC Publishing Group, Chicago IL, 1994
2. Collins COBUILD Dictionary of Idioms, John Sinclair (Founding Editor-in-Chief), HarperCollins Publishers, London, 1995
3. Collins COBUILD English Language Dictionary, John Sinclair (Editor in chief), Collins, London, 1987
4. Merriam-Webster's Collegiate Dictionary, Tenth Edition, Merriam-Webster, 2000

Bibliography

Abeillé, Anne. 1995. The Flexibility of French Idioms: A Representation with Lexicalized Tree Adjoining Grammar. In *Idioms: Structural and Psychological Perspectives*, ed. by M. Everaert, E.-J. van der Linden, A. Schenk, and R. Schreuder, 15–42. Lawrence Erlbaum Associates, Hillsdale, NJ.

Ackerman, Farrell. 1995. Systemic Patterns and Lexical Representations: Analytic Morphological Words. In *Approaches to Hungarian, V: Levels and Structures*, ed. by I. Kenesei, 287–306. Jate, Szeged.

Ackerman, Farrell, and Gert Webelhuth. 1998. *A Theory of Predicates*. CSLI Publications, Stanford.

Anderson, Stephen R. 1992. *A-Morphous Morphology*. Cambridge University Press, Cambridge.

Aronoff, Mark. 1976. *Word Formation in Generative Grammar*. Linguistic Inquiry Monograph 1. MIT Press, Cambridge, MA.

Baayen, Harald. 1992. Quantitative Aspects of Morphological Productivity. In *Yearbook of Morphology 1991*, ed. by G. Booij and J. van Marle, 109–149. Kluwer Academic Publishers, Dordrecht.

Baltin, Mark R. 1987. Heads and Projections. In *Alternative Conceptions of Phrase Structure*, ed. by M. Baltin and E. Kroch, 1–16. University of Chicago Press.

Barlow, Michael, and Suzanne Kemmer. 2000. *Usage-Based Models of Language*. CSLI Publications, Stanford.

Bauer, Laurie. 1990. Be-Heading the Word. *Journal of Linguistics* 26. 1–31.

Beard, Robert. 1990. The Empty Morpheme Entailment. In *Contemporary Morphology*, ed. by W. Dressler et al., Trends in Linguistics - Studies and Monographs 49, 159–169. Mouton de Gruyter, Berlin.

Becker, Thomas. 1990. *Analogie und morphologische Theorie*. Wilhelm Fink Verlag, München.

Bender, Emily, and Dan Flickinger. 1999. Peripheral Constructions and Core Phenomena: Agreement in Tag Questions. In *Lexical and Constructional Aspects of Linguistic Explanation*, ed. by G. Webelhuth, J.-P. Koenig, and A. Kathol, 199–214. CSLI Publications, Stanford.

Bender, Emily M. 2001. *Syntactic Variation and Linguistic Competence: The Case of AAVE Copula Absence*. Stanford University dissertation.

Berko, Jean. 1958. The Child's Learning of English Morphology. *Word* 14. 150–177.

Berry-Rogghe, Godelieve. 1973. The computation of collocations and their relevance in lexical studies. In *The Computer and Literary Studies*, ed. by A. J. Aitken, R. W. Bailey, and N. Hamilton-Smith, 103–112. Edinburgh University Press.

Binnick, Robert I. 1971. Bring and Come. *Linguistic Inquiry* 2. 260–265.

Bird, Steven, and Ewan Klein. 1994. Phonological Analysis in Typed Feature Systems. *Computational Linguistics* 20. 55–90.

Bochner, Harry. 1988. *The Forms of Words: A Theory of Lexical Relationships*. Harvard University dissertation.

Bod, Rens. 1998. *Beyond Grammar: An Experience-Based Theory of Language*. CSLI Publications, Stanford.

Bolozky, Shmuel. 1999. *Measuring Productivity in Word Formation: The Case of Israeli Hebrew*. Brill, Leiden.

Bresnan, Joan. 1982. The Passive in Lexical Theory. In *The Mental Representation of Grammatical Relations*, ed. by Joan Bresnan, 3–86. MIT Press, Cambridge MA.

Bresnan, Joan. 2001. *Lexical-Functional Syntax*. Blackwell, Oxford.

Bresnan, Joan, and Sam A. Mchombo. 1995. The Lexical Integrity Principle: Evidence from Bantu. *NLLT* 13. 181–254.

Brew, Chris. 1995. Stochastic HPSG. In 7th European Conference of the Association for Computational Linguistics, Dublin, Ireland.

Briscoe, Ted, and Ann Copestake. 1999. Lexical Rules in Constraint-Based Grammars. *Computational Linguistics* 25. 487–526.

Bybee, Joan. 1988. Morphology as Lexical Organization. In *Theoretical Morphology: Approaches in Modern Linguistics*, ed. by M. Hammond and M. Noonan, 119–141. Academic Press, San Diego.

Bybee, Joan. 1995. Regular Morphology and the Lexicon. *Language and Cognitive Processes* 10. 425–455.

Bybee, Joan L., and Dan I. Slobin. 1982. Rules and Schemas in the Development and Use of the English Past Tense. *Language* 58. 265–289.

Calder, Jonathan. 1991. Feature-Value Logics: Some Limits on the Role of Defaults. In *Proceedings of the Workshop on Constraint Propagation, Linguistic Description, and Computation*, ed. by M. Rosner, C. J. Rupp, and R. Johnson, IDSIA Working Paper No. 5, Lugano, 20–32. IDSIA, Lugano.

CARPENTER, BOB. 1992. *The Logic of Typed Feature Structures*. Cambridge Tracts in Theoretical Computer Science 32. Cambridge University Press, New York.

CARPENTER, BOB. 1993. Skeptical and Credulous Default Unification with Applications to Templates and Inheritance. In *Inheritance, Defaults and the Lexicon*, ed. by E. J. Briscoe, A. Copestake, and V. de Paiva, 13–37. Cambridge University Press.

CARSTAIRS-MCCARTHY, ANDREW. 1992. Morphology without Word-Internal Constituents: A Review of Stephen R. Anderson's A-Morphous Morphology. In *Yearbook of Morphology 1992*, ed. by G. Booij and J. van Marle, 209–233. Kluwer Academic Publishers, Dordrecht.

CHOMSKY, NOAM. 1980. *Rules and Representations*. Columbia University Press, New York.

CHOMSKY, NOAM. 1985. *Knowledge of Language: Its Nature, Origin, and Use*. Praeger, New York.

CLARK, EVE V. 1993. *The Lexicon in Acquisition*. Cambridge University Press.

CLARK, EVE V., and RUTH A. BERMAN. 1984. Structure and use in the acquisition of word formation. *Language* 60. 542–590.

CLARK, EVE V., and HERBERT H. CLARK. 1979. When Nouns Surface as Verbs. *Language* 55. 767–811.

COOPMANS, PETER, and MARTIN EVERAERT. 1988. The Simplex Structure of Complex Idioms: The Morphological Status of *laten*. In *Morphology and Modularity*, ed. by M. Everaert et al., 75–104. Dordrecht: Foris.

COPESTAKE, ANN. 1992. The Representation of Lexical Semantic Information. Cognitive Science Research Papers 280, University of Sussex.

COPESTAKE, ANN. 1993. The Compleat LKB. University of Cambridge Computer Laboratory Technical Report No. 316.

Copestake, Ann. 1994. Representing Idioms. Presentation at the Copenhagen HPSG Workshop, ms.

Copestake, Ann, Dan Flickinger, Rob Malouf, Susanne Riehemann, and Ivan Sag. 1995. Translation Using Minimal Recursion Semantics. In Proceedings of The 6th International Conference on Theoretical and Methodological Issues in Machine Translation, Leuven.

Copestake, Ann, Dan Flickinger, Ivan Sag, and Carl Pollard. 1999. Minimal Recursion Semantics: An Introduction. Stanford University, http://www-csli.stanford.edu/∼aac/papers/newmrs.ps, ms.

Cronk, Brian C., Susan D. Lima, and Wendy A. Schweigert. 1993. Idioms in Sentences: Effects of Frequency, Literalness, and Familiarity. *Journal of Psycholinguistic Research* 22. 59–82.

Davis, Anthony R. 1996. *Lexical Semantics and Linking in the Hierarchical Lexicon*. Stanford University dissertation.

Davis, Anthony R. 2001. *Linking Types in the Hierarchical Lexicon*. CSLI Publications, Stanford.

Ernst, Thomas. 1981. Grist for the linguistic mill: Idioms and "extra" adjectives. *Journal of Linguistic Research* 1. 51–68.

Everaert, Martin, Erik-Jan van der Linden, André Schenk, and Rob Schreuder (eds.) 1995. *Idioms: Structural and Psychological Perspectives*. Lawrence Erlbaum Associates, Hillsdale, NJ.

Fellbaum, Christiane. 1993. The Determiner in English Idioms. In *Idioms—Processing, Structure, and Interpretation*, ed. by C. Cacciari et al., 271–295. Lawrence Erlbaum Associates, Hillsdale, NJ.

Fillmore, Charles, and Paul Kay. 1997. Construction Grammar. UC Berkeley, http://www.icsi.berkeley.edu/∼kay/bcg/ConGram.html, ms.

Fillmore, Charles, Paul Kay, and Mary O'Connor. 1988. Regularity and Idiomaticity in Grammatical Constructions: The Case of Let Alone. *Language* 64. 501–538.

Firth, John R. 1957. Modes of Meaning. In *Papers in Linguistics*, ed. by J. R. Firth, 190–215. Oxford University Press.

Fleischer, Wolfgang, and Irmhild Barz. 1992. *Wortbildung der deutschen Gegenwartssprache*. Niemeyer, Tübingen.

Flickinger, Daniel. 1987. *Lexical Rules in the Hierarchical Lexicon*. Stanford University dissertation.

Flickinger, Daniel, Carl Pollard, and Thomas Wasow. 1985. Structure-Sharing in Lexical Representation. In ACL 23, 262–267.

Frank, Anette, Tracy Holloway King, Jonas Kuhn, and John Maxwell. 1998. Optimality Theory Style Constraint Ranking in Large-scale LFG Grammars. In Proceedings of the LFG98 Conference, ed. by M. Butt and T. H. King. CSLI Publications, Stanford.

Frauenfelder, Uli H., and Robert Schreuder. 1992. Constraining Psycholinguistic Models of Morphological Processing and Representation: The Role of Productivity. In *Yearbook of Morphology 1992*, ed. by G. Booij and J. van Marle, 165–183. Kluwer Academic Publishers, Dordrecht.

Gazdar, Gerald, Ewan Klein, Geoffrey Pullum, and Ivan Sag. 1985. *Generalized Phrase Structure Grammar*. Harvard University Press, Cambridge MA.

Gelhaus, Hermann. 1977. *Der modale Infinitiv*. Forschungsberichte des Instituts für deutsche Sprache, Mannheim, 35. Verlag Gunther Narr, Tübingen.

Gibbs, Raymond W., and Gayle P. Gonzales. 1985. Syntactic Frozenness in Processing and Remembering Idioms. *Cognition* 20. 243–259.

Gibbs, Raymond W., and Nandini P. Nayak. 1989. Psycholinguistic Studies on the Syntactic Behavior of Idioms. *Cognitive Psychology* 21. 100–138.

Ginzburg, Jonathan, and Ivan A. Sag. 2000. *Interrogative Investigations: The Form, Meaning and Use of English Interrogatives*. CSLI Publications, Stanford.

Goldberg, Adele E. 1995. *Constructions: A Construction Grammar Approach to Argument Structure*. University of Chicago Press, Chicago.

Hantson, A. 1992. Case Assignment and Be-Deletion in Non-Finite Clauses: With Special Reference to (Absolute) Free Adjuncts. *Belgian Journal of Linguistics* 7. 75–94.

Her, One-Soon, Dan Higinbotham, and Joseph Pentheroudakis. 1994. Lexical and Idiomatic Transfer in Machine Translation: An LFG Approach. In *Research in Humanities Computing 3*, ed. by S. Hockey and N. Ide, 200–216. Oxford University Press.

Hoeksema, Jack. 1988. Head-Types in Morpho-Syntax. In *Yearbook of Morphology 1*, ed. by G. Booij and J. van Marle, 123–137. Foris, Dordrecht.

Jackendoff, Ray. 1975. Morphological and Semantic Regularities in the Lexicon. *Language* 51. 639–671.

Jackendoff, Ray S. 1997. *The Architecture of the Language Faculty*. MIT Press, Cambridge MA.

Kaplan, Ronald M., and John T. Maxwell. 1988. An Algorithm for Functional Uncertainty. In Proceedings of COLING 88, Budapest, 297–302.

Kathol, Andreas. 1995. *Linearization-Based German Syntax*. Ohio State University dissertation.

Kathol, Andreas. 1998. Agreement and the Syntax-Morphology Interface in HPSG. In *Studies in Contemporary Phrase Structure Grammar*, ed. by R. Levine and G. Green. Cambridge University Press.

Katz, Jerrold J. 1973. Compositionality, Idiomaticity, and Lexical Substitution. In *A Festschrift for Morris Halle*, ed. by Stephen R. Anderson and Paul Kiparsky, 357–376. Holt, Rinehart, and Winston, New York.

Kay, Paul, and Charles J. Fillmore. 1999. Grammatical Constructions and Linguistic Generalizations: The *What's X doing Y?* Construction. *Language* 75. 1–33.

Keysar, Boaz, and Bridget Bly. 1995. Intuitions of the Transparency of Idioms: Can One Keep a Secret by Spilling the Beans? *Journal of Memory and Language* 34. 89–109.

Kilbury, James. 1993. Strict Inheritance and the Taxonomy of Lexical Types in DATR. Universität Düsseldorf, ms.

Kim, Jong-Bok. 1994. A Constraint-Based Lexical Approach to Korean Verb Inflections. Stanford University, http://www.kyunghee.ac.kr/∼jongbok/papers/final-papers/NEW.ps, ms.

Koenig, Jean-Pierre, and Daniel Jurafsky. 1994. Type Underspecification and On-Line Type Construction in the Lexicon. In Proceedings of WCCFL 13, 270–285.

Koopman, Hilda, and Dominique Sportiche. 1991. The position of subjects. *Lingua* 85. 211–258.

Krenn, Brigitte, and Gregor Erbach. 1994. Idioms and Support Verb Constructions. In *German in Head-Driven Phrase Structure Grammar*, ed. by J. Nerbonne, K. Netter, and C. Pollard, 365–396. CSLI Publications, Stanford.

Krieger, Hans-Ulrich. 1994. Derivation without Lexical Rules. In *Constraints, Language, and Computation*, ed. by C. J. Rupp, M. Rosner, and R. Johnson, 277–313. Academic Press, London.

Krieger, Hans-Ulrich, and John Nerbonne. 1993. Feature-based Inheritance Networks for Computational Lexicons. In *Default Inheritance Within*

Unification-Based Approaches to the Lexicon, ed. by T. Briscoe et al., 90–136. Cambridge University Press.

Langacker, Ronald W. 1987. *Foundations of Cognitive Grammar, Vol I: Theoretical Prerequisites*. Stanford University Press.

Langacker, Ronald W. 1990. *Concept, Image, and Symbol: The Cognitive Basis of Grammar*. Mouton de Gruyter, New York.

Lascarides, Alex, and Ann Copestake. 1999. Default Representation in Constraint-Based Frameworks. *Computational Linguistics* 25. 55–105.

Makkai, Adam. 1972. *Idiom Structure in English*. Mouton, The Hague.

Malouf, Robert. 1994. Noun Incorporation and the Mohawk Lexicon. Stanford University, http://hpsg.stanford.edu/rob/papers/incorp.ps.gz, ms.

Malouf, Robert. 1998. *Mixed Categories in the Hierarchical Lexicon*. Stanford University dissertation.

McCawley, James D. 1981. The Syntax and Semantics of English Relative Clauses. *Lingua* 53. 99–149.

McCawley, James D. 1983. What's with *with*? *Language* 59. 271–287.

McGlone, M. S., S. Glucksberg, and C. Cacciari. 1994. Semantic Productivity and Idiom Comprehension. *Discourse Processes* 17. 167–190.

Melnik, Nurit. 1999. A multiple-inheritance hierarchical representation of Hebrew adjectives of enablement. Presented at the Workshop on Semitic Morphology, University of Illinois at Urbana-Champaign, ms.

Motsch, Wolfgang. 1977. Ein Plädoyer für die Beschreibung von Wortbildungen auf der Grundlage des Lexikons. In *Perspektiven der Wortbildungsforschung. Beiträge zum Wuppertaler Wortbildungskolloquium*, ed. by H. Brekle and D. Kastovsky, 180–202. Bouvier, Bonn.

Motsch, Wolfgang. 1988. On Inactivity, Productivity and Analogy in Derivational Processes. *Linguistische Studien* Reihe A 179. 1–30.

Neumann, Günter. 1997. Applying Explanation-Based Learning to Control and Speeding-up Natural Language Generation. In Proceedings of ACL/EACL-97, Madrid.

Nicolas, Tim. 1995. Semantics of Idiom Modification. In *Idioms: Structural and Psychological Perspectives*, ed. by M. Everaert, E.-J. van der Linden, A. Schenk, and R. Schreuder, 233–252. Lawrence Erlbaum Associates, Hillsdale, NJ.

Nunberg, Geoffrey. 1977. *The Pragmatics of Reference*. CUNY Graduate Center dissertation.

Nunberg, Geoffrey, Ivan A. Sag, and Thomas Wasow. 1994. Idioms. *Language* 70. 491–538.

Olsen, Susan. 1988. Flickzeug vs. abgasarm: Eine Studie zur Analogie in der Wortbildung. In *Semper idem et novus. Festschrift for F. Banta*, ed. by F. G. Gentry, 75–97. Kümmerle, Göppingen.

Orgun, Cemil Orhan. 1994. Monotonic Cyclicity and Optimality Theory. In Proceedings of NELS 24, ed. by M. Gonzàlez, 461–474. GLSA, Amherst.

Pinker, Steven. 1997. Words and Rules in the Human Brain. *Nature* 387-5. 547–548.

Pinker, Steven. 1999. *Words and Rules: The Ingredients of Language*. Basic Books, New York.

Plank, Frans. 1981. *Morphologische (Ir-)Regularitäten. Aspekte der Wortstrukturtheorie*. Studien zur deutschen Grammatik 13. Gunther Narr, Tübingen.

Pollard, Carl, and Ivan A. Sag. 1987. *Information-Based Syntax and Semantics, Volume 1: Fundamentals*. CSLI Lecture Notes Series No. 13. CSLI Publications, Stanford.

Pollard, Carl, and Ivan A. Sag. 1994. *Head-Driven Phrase Structure Grammar*. University of Chicago Press, Chicago and CSLI Publications, Stanford.

Pulman, Stephen G. 1993. The Recognition and Interpretation of Idioms. In *Idioms—Processing, Structure, and Interpretation*, ed. by C. Cacciari et al., 249–270. Lawrence Erlbaum Associates, Hillsdale, NJ.

Quirk, Randolf, Sidney Greenbaum, Geoffrey Leech, and Jan Svartvik. 1985. *A Comprehensive Grammar of the English Language*. Longman, London.

Rainer, Franz. 1988. Towards a Theory of Blocking: The Case of Italian and German Quality Nouns. In *Yearbook of Morphology 1988*, ed. by G. Booij and J. van Marle, 155–185. Dordrecht, Foris.

Riehemann, Susanne. 1993. Word Formation in Lexical Type Hierarchies - A Case Study of *bar*-Adjectives in German. SfS-Report-02-93, University of Tübingen.

Riehemann, Susanne. 1997. Idiomatic Constructions in HPSG. Paper presented at Cornell HPSG Workshop, ms.

Riehemann, Susanne Z. 1998. Type-Based Derivational Morphology. *Journal of Comparative Germanic Linguistics* 2. 49–77.

Riehemann, Susanne Z., and Emily Bender. 1999. Absolute Constructions: On the Distribution of Predicative Idioms. In *Proceedings of WCCFL 18*, ed. by S. Bird, A. Carnie, J.D. Haugen, and P. Norquest. Cascadilla Press, Somerville.

Rosengren, Inger. 1977. *Ein Frequenzwörterbuch der deutschen Zeitungssprache: Die Welt, Süddeutsche Zeitung 2*. CWK Gleerup, Lund.

Sag, Ivan A. 1997. English Relative Clause Constructions. *Journal of Linguistics* 33. 431–483.

Sag, Ivan A. 1999. Explaining the English Auxiliary System. Paper Presented at the Seminar on the Nature of Explanation in Linguistic Theory, UC San Diego, ms.

Sag, Ivan A., and Thomas Wasow. 1999. *Syntactic Theory: A Formal Introduction*. CSLI Publications, Stanford.

Salmons, Joe. 1993. The Structure of the Lexicon: Evidence from German Gender Assignment. *Studies in Language* 17-2. 411–435.

Selkirk, Elisabeth. 1982. *The Syntax of Words*. MIT Press, Cambridge MA.

Shieber, Stuart M. 1986. *An Introduction to Unification-Based Approaches to Grammar*. CSLI Publications, Stanford.

Spencer, Andrew. 1991. *Morphological Theory. An Introduction to Word Structure in Generative Grammar*. Blackwell, Oxford.

Stanley, Richard. 1967. Redundancy Rules in Phonology. *Language* 43. 393–436.

Stump, Gregory. 1985. *The Semantic Variability of Absolute Constructions*. Reidel, Dordrecht.

Stump, Gregory T. 1991. A Paradigm-Based Theory of Morphosemantic Mismatches. *Language* 67. 675–725.

Toman, Jindřich. 1987. *Wortsyntax—Eine Diskussion ausgewählter Probleme deutscher Wortbildung*. Niemeyer, Tübingen.

van Gestel, Frank C. 1992. En-Bloc Insertion. In Proceedings of the Tilburg Idioms Conference, ed. by M. Everaert and E.-J. van der Linden, 75–96.

van Marle, Jaap. 1992. The Relationship between Morphological Productivity and Frequency: a Comment on Baayen's Performance-Oriented Conception of Morphological Productivity. In *Yearbook of Morphology 1991*, ed. by G. Booij and J. van Marle, 151–163. Kluwer Academic Publishers, Dordrecht.

Wasow, Thomas. 1977. Transformations and the Lexicon. In *Formal Syntax*, ed. by P. Culicover, A. Akmajian, and T. Wasow. Academic Press, New York.

Wasow, Thomas, Ivan A. Sag, and Geoffrey Nunberg. 1983. Idioms: An Interim Report. In Proceedings of the XIIIth International Congress of Linguists, CIPL, Tokyo, ed. by S. Hattori and K. Inoue, 102–115.

Webelhuth, Gert. 1994. Idioms, Patterns, and Anthropocentricity. In Proceedings of WCCFL 13, ed. by R. Aranovich et al., 400–415. CSLI Publications, Stanford.

Whitelock, Pete. 1992. Shake-and-bake translation. In Proceedings of COLING-92, 784–789.

Wiese, Richard. 1996. *The Phonology of German*. Oxford University Press.

Zajac, Rémi. 1992. Inheritance and Constraint-Based Grammar Formalisms. *Computational Linguistics* 18. 159–182.

Zeevat, Henk. 1995. Idiomatic Blocking and the Elsewhere Principle. In *Idioms: Structural and Psychological Perspectives*, ed. by M. Everaert, E.-J. van der Linden, A. Schenk, and R. Schreuder, 301–316. Lawrence Erlbaum Associates, Hillsdale, NJ.

Zwicky, Arnold M. 1982. Stranded *to* and Phonological Phrasing in English. *Linguistics* 20. 3–57.

Zwicky, Arnold M. 1985. Heads. *Journal of Linguistics* 21. 1–29.

Zwicky, Arnold M. 1994. Dealing out Meaning: Fundamentals of Syntactic Constructions. In Proceedings of BLS 20, Berkeley, ed. by S. Gahl et al., 611–625.

Zwicky, Arnold M., and Geoffrey K. Pullum. 1986. The Principle of Phonology-Free Syntax: Introductory Remarks. In OSU Working Papers in Linguistics 32, Columbus, OH, 63–91.