University of California

Los Angeles

Learning Form-Meaning Mappings in Presence of Homonymy: a linguistically motivated model of learning inflection

A dissertation submitted in partial satisfaction

of the requirements for the degree

Doctor of Philosophy in Linguistics

by

Katya Pertsova

2007

The dissertation of Katya Pertsova is approved.

Charles Taylor

Colin C. Wilson

Carson T. Schutze

Edward P. Stabler, Committee Chair

University of California, Los Angeles

2007

to my father

Table of Contents

List of Figures

List of Tables

Acknowledgments

This thesis would not be possible without the guidance and inspiration of many of the UCLA faculty. First and foremost, I would like to thank my adviser, Ed Stabler, whose contribution to this work and to my intellectual development and understanding of linguistics in general has been enormous. His classes on learnability and computational linguistics inspired me to pursue this topic, and occasioned many hours of discussion in which he never failed to challenge and to encourage me. His amazing insight, patience and kindness kept me going through the highs and the lows of a graduate student's life.

I am also grateful to other members of my committee: Colin Wilson, Carson Schütze, and Chuck Taylor for stimulating discussion, continuous support, and critical comments.

I am indebted to Donca Steriade for believing in me and encouraging me to enter graduate school, and for her contagious enthusiasm for linguistics. Other faculty that directly or indirectly influenced my work and who I am humbled by are Bruce Hayes, Kie Zuraw, Ed Keenan, and Marcus Kracht.

My parents' love has been unconditional: I owe them everything for their unfailing support in whatever endeavors I undertook.

I've been incredibly lucky to be around the most friendly, fun-loving, and bright community of graduate students, who alone made my years at UCLA pleasurable and worthwhile. My office-mate and dear friend, Sarah VanWagenen, deserves a special thanks for all her help and support over the last few years. For their friendship and companionship, I thank Jeff Heinz, Greg Kobele, Shabnam Shademan, and Dimitrios Nthelitheos. In addition, my years at UCLA would not have been the same without (roughly in chronological order) Joy Elazari,

Mari Nakamura, Asal Sepassi, Leston Buel, Harold Torrence, Adam Albright, Heriberto Avelino, Luka Storto, Jason Riggle, Marcus Smith, Kuniko Nielson, Rebecca Scarborough, Manola Salustri, Julia Berger-Morales, Robert Bowen, Andy Martin, Ben Keil, Mike Pan, Tim and Aki Farnsworth, Christina Kim, Sameer Khan, Asia Furmanska, Ananda Lima and many others (who, hopefully, forgive me if I forgot to mention them).

Last, but not least, thanks to Craig Nishimoto for standing by me through the stress of writing the dissertation, for always being ready to listen and discuss another "scientist's" dilemma, for offering his perspective, for putting all the commas and definite articles into this manuscript, and for his love that gave my life new meaning.

Vita

1976	Born. Moscow. Soviet Union
2001	B.A. Linguistics (with specialization in computing), *summa cum laude*, University of California, Los Angeles
2004	M.A. Linguistics, University of California, Los Angeles

Publications

How Lexical Conservatism Can Lead to Paradigm Gaps, UCLA Working Papers in Phonology 6, 2005

Abstract of the Dissertation

Learning Form-Meaning Mappings in Presence of Homonymy: a linguistically motivated model of learning inflection

by

Katya Pertsova

Doctor of Philosophy in Linguistics

University of California, Los Angeles, 2007

Professor Edward P. Stabler, Chair

In this thesis, I address the issue of learning form-meaning correspondences of inflectional affixes in the presence of homonymy. Homonymy is ubiquitous in all languages despite the fact that it presents a notorious problem for learning and processing. It is a common assumption that patterns of homonymy are restricted in some way and that these restrictions reflect something about the way people learn languages. In this work, I attempt to flesh out this intuition using tools from formal learning modeling.

I show some quantitative evidence that inflectional paradigms have statistical preferences for certain types of non-arbitrary mappings between form and meaning. Namely, one-to-one and "elsewhere" mappings that can be described with defaults are preferred while all other mappings are avoided.

Interestingly, the preferred types of mappings also have a nice learning property: more specifically, there are simple generalization methods that can be used for learning them. The learning model I propose takes advantage of this fact,

although it is still capable of learning 'arbitrary' form-meaning mapping which are empirically attested. Overall, my learner provides a strong bias (rather than a categorical restriction) on the types of patterns it can learn; a bias motivated by the empirical data mentioned above.

The model of learning I propose also predicts intermediate overgeneralization errors and subsequent corrections in the process of language acquisition. It is unique in that, unlike most formal learning models, it relies on a non-monotonic generalization strategy inspired by the blocking proposals in the realm of generative morphological theories.

CHAPTER 1

Introduction

1.1 The Thesis

Children are exposed to a continuous stream of sounds as they experience the world through their perceptual and cognitive systems. Eventually they learn to understand messages encoded by the speech signal and to express similar kinds of messages on their own. One of the central goals of cognitive linguistics is to understand how children gain this ability, or how they acquire language competence.

One way to approach this question is to explore how a computational system might achieve the same competence in a human-like manner, i.e., in a way that captures empirical facts about natural languages and language learning. To be sure, a computational perspective helps us see that there are many ways of learning the same class of languages. However, in trying to understand how human learners do it, it is instructive to pay closer attention to the fine-grain level of empirical generalizations and to the kinds of errors/problems children experience during language acquisition.

One type of fine-grain generalizations are strong statistical tendencies demonstrating that, even when languages don't have categorical limitations on the range of certain options, they might still consistently prefer (to use somewhat vague terms) simple or systematic patterns over more complex and arbitrary patterns.

It is a natural hypothesis that such preferences along with other more categorical universals arise and are maintained in languages because of a particular learning strategy used by human learners (Stabler, forthcoming). In accordance with this hypothesis, paying close attention to preferences and universals exhibited in languages can clue us in to what the human learning mechanism producing these preferences must look like.

In this dissertation, I construct a learning algorithm for learning form-meaning correspondences that is informed by such empirical considerations and that makes further predictions with respect to language acquisition. The domain of my inquiry is the nature of ambiguous form-meaning mappings within inflectional paradigms. Below, I say a few more words about this domain of inquiry and about the main achievements of my dissertation.

There are several reasons for investigating learning lexical meanings of inflectional morphemes. First, learning form-meaning mappings (i.e., learning the lexicon) is fundamental to any theory of language learning since this knowledge is a prerequisite for building meaningful expressions. Second, this domain of inquiry is relatively understudied, especially below the word level (for some work in this direction see, however, Albro (1997); Carlson (2005); Adger (2006)).

Anyone who contemplates lexical learning for a few minutes will realize that ambiguity (or deviations from the one-to-one mapping between form and meaning) present a problem. Commonly, it is implicitly assumed that patterns of ambiguity (especially homonymy/syncretism) in inflection are connected to the properties of the human acquisition device (Williams, 1994; Wunderlich, 2004, and others). My work is an attempt to flesh out this assumption into a formal learning model. In pursuing this goal, I adopt the hypothesis that the learning of form-meaning mappings involves default reasoning (introduced in the next sec-

tion). Roughly stated, default reasoning involves default rules that apply only when other rules fail to apply, as in the statements: *if* X *then* Y *else if* Z *then* W *else* Q. This view leads me to define precisely which form-meaning mappings are describable with defaults (without positing homonymy), and which are not. Given this definition, and my particular definition of homonymy that takes the learner's point of view into consideration (see next section), I address the question of what types of form-meaning mappings are empirically attested in languages and to what degree. Based on a sample of verbal agreement paradigms from 30 genetically diverse languages, I find that

(1) a. Non-homonymous mappings predominate in these paradigms

b. Among homonymous mappings those that can be described with defaults are by far the most dominant.

In the end, I propose a formal learner that can handle any attested form-meaning mapping, but that matches the discovered statistical tendencies by generalizing in such a way that non-homonymous mappings are the easiest to learn, followed by default mappings, followed by what I call "overlapping mappings" (i.e., mappings not describable by default reasoning). Additionally, my learner learns in the presence of irrelevant features (i.e., it does not know *a priori* which semantic contrasts out of all possible contrasts are grammaticalized in the target language), and it predicts overgeneralizations at intermediate learning stages followed by subsequent corrections – a pattern of behavior also characteristic of human learners (Marcus et al., 1992; Strauss and Stavy, 1982; Marchman et al., 1997).

1.2 Non-monotonicity

The learner I propose in this thesis is unique because it relies on a non-monotonic[1] learning strategy unlike the overwhelming majority of the formal learning models. Monotonicity is preferred in formal learning modeling because it allows the learners to generalize in a conservative fashion (without making errors) and keeps the learning strategies and the proofs about them simple since the truth is preserved at every intermediate step.

However, these advantages do not by themselves constitute a reason for believing that human learners are monotonic. In fact, the overgeneralization errors reported by many researchers on language acquisition are more consistent with the non-monotonic picture of learning.

Besides, non-monotonic reasoning appears to be natural and commonplace in making inferences and decisions in the face of incomplete or changing information. Such reasoning usually involves relying on a general rule of thumb that captures *typical* cases and that has exceptions. For instance, consider the following example of non-monotonic reasoning from the realm of language processing (from Antoniou, 1997). Suppose we are reading a text that begins like this:

Smith entered the office of his boss. He was nervous.

At this point, most readers would assume that the pronoun *he* refers to Smith. But the immediately following sentence (below) is inconsistent with this assumption, and so will most likely lead the readers to revise their current hypothesis:

After all, he didn't want to lose his best employee.

[1] A non-monotonic learner is a learner whose intermediate hypotheses don't grow monotonically. That is, such a learner may converge on a language that is smaller than the learner's preceding hypotheses. In simpler terms, a non-monotonic learner may overgenerate at intermediate stages and later correct such overgeneralizations.

Perhaps in the ideal world, we would have enough information (or we would wait until we have enough information) to make our decisions, including a decision about what "he" refers to in the above text. But in reality, we often rely on rules of thumb that work most of the time, but that ultimately have exceptions. A learner only beginning to learn a language is precisely in the situation in which he or she has quite impoverished and incomplete information, and so the use of non-monotonic reasoning is only natural (while of course not necessary, especially if the language is restricted in such a way that it's possible to generalize and never be wrong[2]).

While formal models avoid non-monotonic reasoning, traditional descriptive models of language relying on non-monotonic representations (and often implicitly assuming non-monotonic learning) are quite common in linguistics, cf. "the Elsewhere Condition" (Panini, Kiparsky (1973)), the "Subset Principle" (Halle, 1997), the blocking rules of Aronoff (1976), aspects of the Optimality Theory (Prince and Smolensky, 1993), etc. I will lump all such proposals under the general rubric of *blocking proposals*. The essence of the blocking proposals is that the grammar involves competition among different rules (or principles), and a way to determine which rules "win" the competition in particular cases. The winning rules can "block" the application of other rules which are then said to have "default" status applying only as a last resort in a particular sub-domain. (Notice, that there might be several default rules in a system, as they can be nested in each other or disjoint[3].)

The most prominent arguments for descriptive systems involving defaults are based on economy considerations. In section 2.1.2.3 (chapter 2) I show that such arguments are not convincing, especially in the domain of inflection. The learning

[2]For an example of such a learner in the domain of learning phonotactics see Heinz (2007).
[3]For more examples of cases with several defaults see figure 3.1 on page 53.

model I present here, on the other hand, provides a stronger reason for adopting such representations - it shows that a learner biased to use default reasoning (and producing grammars with blocking) gives us a certain fit with frequencies of different form-meaning correspondences found in inflectional paradigms. Moreover, this learner makes testable predictions with regard to language acquisition and language change, which could potentially provide further support for this model (or to illuminate ways in which it can be improved).

1.3 Patterns of inflectional homonymy: definitions

In this section, I go over some important definitions related to the central notion of this thesis, homonymy, which presents a problem for learning form-meaning mappings.

But let me first clarify some terms that are used in the subsequent definitions. I use the term *morph* to refer to the phonological realization of a *morpheme* which is in turn conceived of as a lexical unit having several components: a phonological component (the morph), and the semantico-syntactic components specifying the distribution of this morph in the language. (See next chapter for the discussion of alternative conceptions of morphemes and morphological structure in general).

Morphology abounds with cases in which a single morpheme is used in several different ways (in linguistic representations this happens when it occupies more than one cell in a paradigm). Throughout this dissertation I will refer to this phenomenon as *form identity*.

Certain instances of form identity are due to *homonymy* (or semantic ambiguity), while others are due to the fact that some inflectional contrasts are *irrelevant* in particular environments (as exemplified shortly). In morphology,

the term "homonymy" is used in many different ways. I will use it in a somewhat non-standard fashion relying on the neutral notion of "distribution" rather than the notion of "lexical meaning" that imports various assumptions about the structure of the lexicon.

Normally, one would say that two morphemes are homonymous if they sound the same but have different lexical meanings. This assumes that we already know which morphs are distinct despite having the same form and what their lexical meanings are. However, since the learner does not initially know which same-sounding morphs are distinct, the standard definition above is not suitable for our purposes. The only thing that the learner has access to is the distribution of morphs. There is syntactic distribution (which other morphs a given morph can occur with, in what order it occurs, etc), and semantic distribution (what semantic properties must be satisfied for a given morph to be licenced). Focusing mainly on the latter notion of distribution, we can observe that if such distribution can be correctly described with a *single set of necessary and sufficient semantic features*, then it is always possible to equate this set to the morph's content or "meaning"[4]). Such a morph should not have a status of a homophone under any standard theory since it can be assigned a single lexical meaning.

Otherwise, if a morph's semantic distribution cannot be described with a single set of necessary and sufficient features, something special has to be done to capture its meaning, e.g., positing defaults and blocking, or positing separate homonymous lexical entries, or allowing conjunction of feature sets, etc. I will restrict the term *homophone* (or homonym) for this latter scenario only. So, a homophone is a morph that can be used in several different ways and that meets

[4]The word "meaning" here is used to refer to the internal lexical representations in the speakers' mental lexicon, rather that the externalist notion of meaning argued for in the philosophical literature.

the following definition:

(2) A morph is a *homophone* if its distribution cannot be described in terms of a single necessary and sufficient set of semantic values (and this is not due to free variation).

For example, on this definition *are*, the present tense form of the verb *to be*, is a homophone since there is no single set of semantic values that would accurately describe its distribution. The set [BE, pres.tense, indicative] is necessary but not sufficient since these semantic values are also compatible with forms *is* and *am*.[5]

An example of form identity that is not due to homonymy, but to irrelevant contrasts, is the use of the French plural determiner *les*. One would typically say that *les* could be used either with masculine or feminine nouns because gender is irrelevant in the plural, and not because there are two different homonymous determiners *les*. This intuition is usually captured with the notion of feature underspecification (discussed in section 2.2.2).

In the learning chapter, I will also use the term "homonymous lexical entries" for the situation when the learner has already acquired some portion of the lexicon and in this lexicon several distinct lexical entries have the same pronunciation.

The definitions presented here are crucial for understanding other distinctions and terms that will be introduced as we go along.

[5]It is possible to describe the distribution of *are* with a single lexical entry that has a default status provided some assumptions about how such representations should be interpreted. Alternatively, one can posit several different lexical entries for *are*. At this point I am not concerned with the differences between such accounts; I'm merely illustrating my use of the term "homonym".

1.4 Assumptions about prior knowledge

In this section, I present the basic assumptions regarding my learner's capacities and prior knowledge. Some of these capacities/knowledge are hypothesized to be innate, while others are attributed to previously acquired information. Several of the assumptions discussed below present simplifications which we would eventually like to relax, but which are useful in tackling a complex problem with many interacting factors.

First, I assume that there is a finite set of universal distinctions that can be encoded by means of inflection. All languages draw from this universal set, but they differ in what distinctions they end up encoding. Also, I assume that languages are compositional; that is, the meanings of larger structures are determined from the meanings of smaller structures together with the rules of composition.

Second, I endow my learner with some prior knowledge based on the assumption that when acquiring meanings of inflectional morphemes, children do not start from the "blank slate." We have reasons to believe that by the time they begin acquiring morphology, they already know quite a lot about the phonological forms of their language and they have already developed some conceptual representations. That is, I assume that the learner already comes to the task of learning morphology with some knowledge about basic units of form and the ability to "perceive" meaning. In particular, I assume that strings of phonemes corresponding to phonological realizations of inflectional morphemes have already been identified. Additionally, I assume that the learner has the ability to perceive and infer from the environment (I use the term 'environment' in the broadest sense possible) the semantic values of the universal inflectional distinctions. Both of these (obviously, idealized) assumptions are discussed at greater length below.

The first assumption finds some support in the fact that it is in principle possible to discover many morphs without any semantic information. Roughly speaking, this can be done by looking for a minimal number of phonological chunks that repeatedly co-occur in the speech stream and that obey certain prosodic (and other linguistic) constraints. There are several computational algorithms that more or less rely on this idea to find morpheme boundaries in a continuous text of phonemes or graphemes (de Marcken, 1996; Brent, 1999; Goldsmith, 2001; Baroni, 2003). Most of these algorithms are based on purely statistical and distributional information, but incorporating some linguistic biases into such models significantly improves their performance (Cambell and Yang, 2005).

Infant studies also lend support to the idea that humans are able to use statistical information to "jump start" the segmentation process, and as they learn more about the input, other cues to word and morpheme boundaries such as stress, intonation and phonotactics begin to play an increasingly important role. For instance, we know that young infants can track transitional probabilities of syllables even after very brief exposure to the training data (Saffran et al., 1996; Aslin et al., 1998). Nine month old English speaking infants are already sensitive to actual prefixes of their language, but not yet to the suffixes (Santelmann et al., 2003). Several studies show infants' sensitivity to stress and phonotactics when these are used to mark morpheme boundaries (Mattys et al., 1999; Johnson and Jusczyk, 2001; Thiessen and Saffran, 2007).

The second assumption I mentioned has to do with semantic representations. I assume that at the onset of learning all possible semantic features that could potentially be expressed by inflectional morphemes are available to the learner, and that learners are capable of determining values of these features based on perceptual information, cognitive inferences about speakers' intentions and even

semantic information (see below and page 32 for a discussion of exceptions to this assumption.) The question of how exactly are the contrasts perceived and/or inferred from the environment is still an open question in the domain of psychology, and I don't have much to say about it.

Recall that I assume that in the process of learning, the learners come to figure out which of the universally possible contrasts are encoded in their language and which are *irrelevant*.

There are other domains in language acquisition where there is evidence that children initially pay attention to lots of contrasts, but gradually stop paying attention to those contrasts that do not prove to be useful. For example, when it comes to speech perception, 6 month olds can distinguish practically any non-native phonetic contrast, but by 12 months of age this ability declines and infants reliably discriminate only those phonetic contrasts that are phonemic in their language (see review by Werker (1989)). Similarly, in the domain of word learning, it has been shown that 13 and 18 month olds generalize a learned object name to new instances based on overall similarity across many dimensions (Smith et al., 1999). But by age 2, children start showing systematic biases, attending to specific dimensions for different types of objects – *shape* for the artifact-like things, *material* for substances, *colors* for foods (Imai and Gentner, 1999; Booth and Waxman, 2002; Jones and Smith, 2002).

It is worth noting that some inflectional morphemes express meanings that in principle cannot be learned from the environment, such as inflection classes, gender of inanimate nouns, some case marking, etc. These features mark either syntactic or arbitrary relationships, and they have to be learned from the distributional or syntactic information. Learning how such features are mapped to morphs is largely outside the scope of this thesis (see, however, discussion

at the end of chapter 5 for some remarks about possible directions for learning inflectional classes).

Provided the two assumptions above, the first rough characterization of the learning problem I tackle can be stated as follows: given a string of inflectional morphs uttered in a particular situation that can be described in terms of a complete assignment of all universal features to their values, the learner has to determine which of the features are relevant, and how they match up with the individual morphs.

To give a more concrete example, imagine that upon hearing a word *"elephant-s,"* the child can infer from the situation that this word refers to the big grey animals with trunks, that there are more than one of them, that they are "animate", they are "definite" (the particular elephants standing over there), they are located in front of the child, they are present now, they are relatively far away, etc. Given all this (and other similar kinds of) information, the child has to figure out that *-s* (and not *elephant*) encodes the property "plural" (and not definiteness, location, animacy, etc). Later on, when a child experiences the use of *-s* to mark possession (as in *an elephant's trunk*), she would also have to correctly resolve the ambiguity and be able to detect that this time *-s* is used in a very different way and does not indicate the property "plural".

1.5 Thesis Outline

The general structure of this thesis is as follows. In the next chapter, I will discuss some of the basic concepts pertaining to the structure of lexicons. I will also introduce a first intuitive proposal about how lexical meanings might be learned and show how this proposal, in its simplest formulation, fails to deal

with homonymy. Nevertheless, the basic idea behind this proposal will play an important role in the learning algorithms proposed later.

In chapter 3, I concentrate on the theoretical issues surrounding homonymy and syncretism in inflectional paradigms. Here is where I define the notions of "elsewhere" and "overlapping" homonymy, and formulate the empirical hypotheses with respect to frequency of different patterns of form-meaning mapping. These hypotheses are evaluated against typological data and against calculations of chance frequencies in chapter 4. Finally, chapter 5 is devoted to the learning model. This chapter begins with some general discussion of adopted assumptions and definitions related to formal learning theory. I then proceed to present three learning algorithms building up to the final General Homonymy learner. Each new algorithm covers more empirical ground, and builds on the previous simpler algorithm. A thorough understanding of this chapter may require familiarity with formal notation. However, such knowledge is not required for getting the grasp of the basic ideas.

CHAPTER 2

Lexicon and cross-situational learning

This chapter lays a foundation for the rest of this dissertation. Here I describe general assumptions about the organization of the lexicon and introduce some terminology and key concepts that are used throughout the thesis.

I begin by providing background on certain common assumptions about lexical representations. In the second half of this chapter, I discuss a "cross-situational" approach to acquiring lexical meanings that gives the reader a first glimpse at a general learning strategy which forms the backbone for the formal work presented in chapter 5.

2.1 The nature of the morphological lexicon

2.1.1 Lexical units

The first question that arises when one talks about lexical learning is what are the appropriate lexical units in speakers' mental lexicon? This dissertation rests on the assumption that regular inflectional markers, such as affixes, are among such atomic lexical units. This assumption is not without controversy, as some researchers hold a view that speakers don't decompose words into morphemes but rather store them as a whole (Butterworth, 1983; Seidenberg and McClelland, 1989; Gonnerman, 1999). In such models, morphemes are discussed as epiphe-

nomenal objects that amount to semantic and acoustic/orthographic similarities among words, as opposed to abstract units that have their own lexical representations. Morphological productivity is accounted for by appealing to analogy or to rules derived by mechanisms of general pattern extraction based on a subset of words that are similar in some relevant respect.

The opposition between the two views (storing words as decomposed or as a whole) might not be as drastic as it appears at the first glance. Once one specifies precisely what the rules of pattern extraction are and how similarity of words can be used to compute the relationships between overlaps in form and overlaps in meaning, I believe that the two points of view will be very difficult to distinguish from each other on the basis of their predictions about what's grammatical. However, they do make somewhat different processing predictions.

Some of the latest experiments testing these predictions (using the lexical priming paradigm) support the morphemic point of view, where morphemes rather than words are the atomic units stored in the lexicon.[1] Priming is based on the idea that accessing a lexical representation in the mental lexicon will facilitate subsequent access of the same lexical representation as well as of other semantically or formally similar representations. Proponents of whole word storage maintain that morphological priming effects are reducible to the sum of semantic and formal priming. However, it has been established that in certain experimental conditions, when the prime and the target are separated by several other words (long-lag priming), the semantic and formal priming do not obtain, i.e. *jump* does not prime *hop*, and *car* does not prime *card*. In such conditions, morphological priming effects persist (*sings* continues to prime *sing* and *happiness* continues to prime *shyness*) suggesting that morphemic representations can

[1]This does not mean that whole words or even whole phrases cannot be stored as a whole if they cannot be analyzed compositionally.

prime each other independently from phonological and semantic representations (Bentin and Feldman, 1990; VanWagenen, 2005).

Stockall and Marantz (2006) report results from a MEG study that show reactivation effects even for the regular-irregular verb pairs whose overlap in form is rather minimal (e.g., *teach* - *taught*). They also mention a study on Finnish by Jarvikivi and Niemi who showed that monomorphemic words (like the singular noun *sormi* "finger") can be primed by a bound stem allomorph which is not a real word of Finnish (*sorme* from *sormesta* "from finger"). At the same time, phonologically matched pseudo-words such as *sorma* do not lead to priming. This experiment suggests that both roots and stems have their own lexical representations. The results are not easily explained by the whole word storage model, since the two primes - *sorme* and *sorma* - overlap with the target in form and meaning (or the lack thereof) to the same extent. The only difference between these pseudo-words is that one is a possible bound stem while the other is not.

The view that morphemes are lexical units is also more intuitive given a natural hypothesis about how lexical knowledge might be acquired. Consider a problem a child faces when trying to parse the continuous stream of speech and make sense of it. We have reasons to believe that even before children understand simple sentences, they have already begun to segment speech into discrete units that later on will be mapped onto conceptual structures. Our best models of segmentation so far are mainly based on distributional evidence (see section 1.4) and draw no principled distinction between words and morphemes. If anything, the criteria they use for finding boundaries in phonological strings leads to the discovery of morphological units, not of words (de Marcken, 1996; Goldsmith,

2001; Baroni, 2003).[2] Likewise, the distinction between words and morphemes, although appearing intuitive to us, is notoriously hard to draw on theoretical grounds (Williams and DiScullio, 1987). Given that whole-word theories of morphological organization make a distinction between words and morphemes, where the former are units of meaning listed in the mental lexicon and the latter are epiphenomenal objects, one might ask how a child would arrive at this rather shaky distinction in order to store words but not morphemes? For instance, if a child is learning a fairly well-behaved agglutinative language, what would prevent her from using general learning strategies for segmentation and association of forms with meanings to posit morphemic lexical entries? Such learning strategies are necessary in any case for discovering atomic units to be stored in the lexicon (whatever those units might be).

Another anti-morphemic view is maintained by the proponents of the Word and Paradigm tradition who claim that inflectional marking is achieved by means of transformations applied to the stem (Zwicky (1985); Anderson (1992); Stump (2001) and others). In these models, stems or "bases" are listed in the lexicon proper while inflectional rules are part of a separate grammatical component consisting of rules that specify how inflectional features should be realized. These models are motivated by the fact that inflectional systems can contain non-concatenative and irregular means of grammatical marking. On the other hand, fully morphemic approaches make no distinction between stems and other morphemes; they are all conceived of as "pieces" that are combined together either in the lexicon itself (Lieber, 1992) or in the syntax (Marantz, 1997). Lieber proposes that non-concatenative irregular patterns can be dealt with by means of autosegmental and prosodic phonology such as floating features, etc. Marantz and the Distributed Morphology (DM) tradition assume a special battery of readjust-

[2]De Marcken's model produces a hierarchy of units including phrases, word and morphemes.

ment rules that apply post-syntactically to handle irregular morphology (some irregularity is also handled at the lexical insertion). Finally, there are dual-rule models where morphemic representations are assumed only for regular and concatenative morphology, while all other words are not decomposable but stored as a whole (Pinker, 1991; Marcus, 1995; Clahsen, 1999).

These alternatives remain hotly debated. I will avoid this debate by focusing my attention on concatenative and regular patterns of affixation. For this subtype of inflection any of the above mentioned approaches assume that there is some association between the phonological realizations of grammatical distinctions (morphs) and the features or representations they are associated with (whether we want to call this association a "rule" that applies to stems, or a lexical item which directly encodes both the phonological and the semantic components of the morpheme). I believe that the same largely holds for non-concatenative inflection if one does not restrict morphs to a contiguous string of phones.[3]

Looking at the concatenative inflectional patterns is just a first step in understanding how form-meaning mappings are learned. We have to start somewhere, and I prefer to start with simple cases before proceeding to more complex ones. This endeavor is not invaluable especially given the fact that concatenative inflection seems to predominate cross-linguistically. For instance, Greenberg (1963) observes that most languages in his sample use affixation to mark inflectional contrasts. The predominance of affixal inflection is also true for the sample of 30 languages I will discuss in this thesis (however, the languages in my sample were not selected completely randomly but with an eye towards systems with

[3]As I see it, the main difference between morphemic and Word and Paradigm approaches is not in how they instanciate the relationship between forms and meanings, but in the difference of the status attributed to the stems. In the Word and Paradigm approach one of the stems per lexeme has a special stutus of a "base" from which all other forms are derived, including other related stems. No such difference exists in morphemic approaches: all morphs, including roots and stems, combine with each other in the same way.

syncretism).

2.1.2 Minimality and non-redundancy

Besides the fact that lexicons contain morphemic representations, they are also often assumed to be somehow *minimal* and/or *non-redundant*. The notion of minimality has been one of the central notions in the generative linguistics, albeit a difficult one to define precisely.[4]

There are two different kinds of minimality or economy proposals in the literature. First, there are proposals that certain structures are avoided because they are non-minimal. Second, there are proposals that certain descriptions or representations of structures are avoided because they are non-minimal. An example of the first kind of proposal is the conjecture that perfect synonymy is dispreferred for reasons of economy. A lexicon with abundant synonymy or free variation not only would have more lexical entries than a lexicon without free variation, but it would also generate more strings.

An example of the second kind of minimality has already been alluded to in this chapter: morphological models that assume full decomposition are more economical in the sense that they posit fewer lexical entries than the whole-word models, but both are intended to generate exactly the same strings. The

[4]One of the difficulties is that what is minimal for one aspect of language is not necessarily minimal for another aspect. For example, Plank (1986) observes that agglutinative or separatist inflectional systems (where every inflectional feature is realized by a separate morph) allow for shorter lexicons, but result in longer strings and hence require more effort for the production system. The cumulative inflection (several features realized by the same morph) lead to longer lexicons, but result in shorter strings. To see this, consider the fact that given two features with three values (6 values all together), there are $3^2 = 9$ distinctions that can be made. A language that makes all these distinction via cumulative affixes will need 9 morphemes, whereas a language in which these distinctions are made by combining separate morphs will only need 6 morphemes (one for each feature value). But, the first language will realize the two features using just one morph, while the second language will have to use two morphs for the same purpose.

hybrid models of the lexicon (which assume both whole word and decomposed representations for some words) choose to economize on the processing time and effort rather than on the size of the lexicon (cf. Augmented Addressed Morphology, Caramazza et al. (1988)) or Morphological Race Model, Frauenfelder and Schreuder (1992)).[5]

The idea that language users and analysts should prefer shorter descriptions was already present in the SPE rule model of Chomsky and Halle (1968). Formal notions of this idea were developed in the domain of information theory and gave rise to the so called "minimum description length" approach (Wallace and Boulton, 1968; Rissanen, 1978). The basic principle of this approach rests on the hypothesis that *all else being equal* shorter descriptions are simpler and therefore more likely.

In this section, I consider three assumptions about minimizing descriptions in the domain of morphological lexicons: exclusion of irrelevant features from lexical representations, the use of null morphs, and the use of blocking rules. These assumptions are motivated by considerations of storage economy and are often adopted as constraints on the descriptive apparatus (the lexicon). As a side note, although such restrictions on grammars seem prima facie reasonable, they are difficult to test or confirm empirically. This is because the predictions they make concern rather subtle facts about processing rather than facts about grammaticality. However, as I show in this thesis, some of the proposals above can be restated as proposals about the learning algorithm, which does make testable predictions, namely predictions about overgeneralization errors in the process of

[5]In such models, memory recall and morphological analysis run in parallel. The memory recall is faster and more efficient for high frequency words, while the morphological analysis is faster and more efficient for low frequency words (some of which lack whole-word representations all together). Since both of the routines apply in parallel until one of them succeeds, this ensures that the most efficient strategy is applied in each case.

language acquisition and about statistical frequencies of patterns that are harder to learn (and harder to describe succinctly within a particular framework).

2.1.2.1 Exclusion of irrelevant features

It is a common (and mostly implicit) assumption that lexicons *do not* include irrelevant features in the representations of morpheme meanings. Irrelevant features are not overtly marked either in the language as a whole or in certain contexts (see section 2.2.2). For example, we don't see morphological analyses of the following sort.

(1) Lexical entries for the English plural morpheme -*s*:

a. -*s*: [+pl,+anim,+fem]

b. -*s*: [+pl,+anim,−fem]

c. -*s*: [+pl,−anim]

Although the above lexicon correctly predicts how the plural morpheme is used, an alternative and generatively equivalent lexicon with a single lexical entry -*s*:[+pl] is more minimal. If lexicons always specified irrelevant features for every morpheme, they would contain an enormous amount of redundant homonymy. In the worst case, every morph would have as many meanings as there are different situations in which it could be used, which would defeat any usefulness of morphological analysis. Moreover, this kind of redundancy would fail to encode the generalization that phonologically similar inflectional morphemes are also usually semantically similar.

The fact that lexicons do not include irrelevant features is usually stated as a requirement to use *feature underspecification* in lexical representations whenever

possible. Bierwisch (2006) puts it this way: "The quest for economy ... leads to the assumption that lexical representations are subject to underspecification, such that lexical entries respect in one way or the other the conditions that make predictable specifications follow from more general rules or principles." In this work, I will also assume that morphemic representations are maximally underspecified (in the "strict" sense of underspecification which I explain in section 2.2.2). This requirement is built into the formal description of the target lexicons for the learning algorithm in chapter 5.

2.1.2.2 Null morphs

Positing null morphs to describe non-overt realization of meaning also helps us to avoid positing redundant homonymy. To see this, consider the following inflected words from Russian.

(2) stran-a ("country", nom.sg.)
ruk-a ("arm", nom.sg.)
stran ("country", gen.pl.)
ruk ("arm", gen.pl.)

Taking this mini-set of words in isolation, we have several choices in how to assign meanings to the individual morphs in the example. If this were a problem set for Linguistics 1, most students would quickly determine that the meaning of the suffix *-a* is [nom.sg]. As for the other morphs, there are several options. One option is to assume that there is a null (silent) morph that expresses the meaning [gen.pl.]. This morph attaches to the stems *stran-* ("country") and *ruk-* ("arm") in the same way as the suffix *-a*. Another option is to posit two separate lexical entries for each of the roots. For example, the root *stran* could be associated with

two meanings "country" and "country, gen.pl.". This means that thousands of other words like "country" and "arm" would also have two homonymous roots. It is obvious that the first option - positing a single null morph - is more economical and avoids unnecessary redundancy in lexical entries.[6]

In this thesis I will take for granted the idea that null morphs are part of the morphological vocabulary since they are useful in succinctly describing data like the Russian example above. However, I will not address the question of how they may be discovered and learned, instead I will assume that they are supplied by the segmentor (see, however, some preliminary ideas for the problem of learning null morphemes in chapter 5, section 5.2.3).

2.1.2.3 Blocking and minimality

Another descriptive tool that arguably has a minimizing effect on the size of lexical representations is the assumption of *blocking* mentioned in chapter 1 in connection to default reasoning. One of the most wide-spread uses of blocking is to capture irregular morphology. For example, the English past tense is often analyzed by specific rules or specific lexical items for the irregular verbs (such as *taught, spent, sang*) and a general default rule for the regular *-ed* affixation (*jumped, walked, yelled*). The irregular verbs are said to "block" the application of the regular *-ed* affixation. The use of the blocking principle can be viewed as a filter on the expressions generated by the lexicon. Those expressions that are not

[6]In some theories in which features are monovalent, the unmarked values are assumed by default and do not have to be specified in lexical representations. On this view of features, non-overt realization of meaning can be easily explained without positing null morphs or redundant homonymy, but only if such non-overt realization always coincided with the expression of unmarked values. Although languages do show a correlation between zero-marking and semantic non-markedness (cf. Jakobson, 1939), it is at best only a tendency. In the Russian example above, the feature values "genitive" and "plural" are not the unmarked values for the categories of case and number. Therefore we can't assume that these features would be provided as default features in the absence of an overt marker.

"filtered out" or blocked are grammatical, while all others are ungrammatical. In other words, there are two components to the grammar - a lexicon which is allowed to overgenerate, and a blocking principle (or blocking rules) which rule out overgenerated expressions. (This view does not commit us to a processing model in which filtering is a second stage that follows a first stage of overgeneration.) The blocking principle can be formulated in many ways, depending on the empirical facts. The most common way used in linguistics is to say that more specific rules or lexical items block the more general ones (although see discussion in section 5.5.2 of the empirical vacuousness of this principle).

The two-component grammar (lexicon with defaults + blocking principle) is often shorter than an alternative description consisting of a single lexicon in which lexical representations alone are sufficient for generating only grammatical expressions. For example, in the case of the English past tense, the lexical entry for *-ed* in the description without the blocking principle would have to include a list of all regular verbs with which *-ed* can be used (since the membership in either regular or irregular class is largely arbitrary).[7] This of course requires listing thousands of stems because the regular verbs constitute a majority of English verbs.[8]

On the other hand, in the description involving a blocking principle, we only need to list irregular verbs (either as contextual restrictions on irregular rules or as independent lexical items). The *-ed* suffix is then said to have an elsewhere distribution (i.e., during the insertion process it will apply only to those stems

[7] I assume that lexical entries not only specify the semantic content or meaning of morphemes, but also contextual information encompassing any idiosyncratic facts about how the morpheme in question is to be used.

[8] Another alternative would be to assume that the contextual specification of the morpheme *-ed* was something like "is NOT used with sing, teach, rise, etc". However, such negative specifications of lexical items are viewed as unacceptable by some morphologists (e.g., Carstairs, 1998). Additionally such a lexicon will still be less minimal than the lexicon in which *-ed* is simply stipulated as a default morpheme, since it would mention the irregular verbs twice.

that are not listed as irregular).

When it comes to inflectional paradigms, the blocking principle (instantiated as the Subset Condition) together with underspecification is often used to describe certain patterns of homonymy.[9] In this domain, however, the savings offered by the use of blocking are much less significant given that inflectional paradigms are usually small in size to begin with.

Additionally, even if the blocking accounts are somewhat more minimal, they achieve this minimality by shifting the complexity from the lexicon to the processor. For instance, consider two alternative accounts of the present tense paradigm of the English verb "to be".

(3) Two alternative descriptions of the present tense of "to be"

a. With no blocking

am [BE, pres., 1p., sg.]

are [BE, pres., 2p., sg.]

is [BE, pres., 3p., sg.]

are [BE, pres., pl.]

b. With blocking

am [BE, pres., 1p., sg.]

is [BE, pres., 3p., sg.]

are [BE, pres.]

Subset Principle: more specific items block more general ones.

The second account might be just a tiny bit more economical than the first one in the number of lexical entries, but it involves an additional blocking component

[9]Blocking proposals also have an effect of ruling out free variation, which appears to be rare in inflection, although not non-existent.

which introduces an extra reasoning step during the generation/production of phonological forms. Suppose we're trying to generate the phonological realization of [BE, pres.,1p.,sg.]. Given the first account, we just look up which lexical item is consistent with this meaning. Given the second account, we do the same thing except this leads to competition between *is* and *are*, and we need to apply the blocking principle to resolve it. In other words, it is not obvious that one of the accounts above is more minimal than the other. In general, the minimality argument does not provide a convincing motivation for preferring the blocking accounts of type (b) above to the generatively equivalent accounts of type (a).[10]

Nevertheless, I will adopt the blocking types of descriptions as targets for my learners, but for reasons other than minimality. More specifically, adopting such non-monotonic representations will allow my learners to use a natural generalization strategy and a simple way of correcting overgeneralizations, at the same time as accounting for the statistical tendencies found in patterns of form-meaning mappings (see the next two chapters).

2.1.3 Interim summary

To summarise the discussion so far, a lexicon is a theoretical device we posit to account for our conviction that speakers must have some mental repository of

[10]Sometimes, there are other arguments suggested in the literature for preferring descriptive accounts involving blocking. For instance, it is claimed that such an account predicts how paradigm gaps should be filled in paradigms with defaults (Halle and Marantz, 1994). However, it is easy to see that any generalization that can be expressed in an account of type (b) can also be expressed in an account of type (a) since there is a direct translation from one formalism to the other. In particular, if one makes an additional assumption (and it really is an additional assumption in disguise) that paradigm gaps should be filled by defaults, then the same assumption can be made in the alternative account, except we would have to explicitly specify the properties of morphemes that can be extended to cover paradigm gaps. Besides this conceptual point, there is also lack of conclusive empirical data showing that paradigm gaps indeed tend to become filled by forms that can be independently shown to have a default status.)

associations between units of form and units of meaning. This repository is part of the grammar which allows speakers to generate and understand expressions of their language. I assume that inflectional affixes are among the lexical units stored in the lexicon. The minimal amount of information that a morphological entry must encode includes its phonological form and the semantic content which specifies the distribution of this form with respect to semantic environments (it may also include contextual and syntactic restrictions on its distribution). Additionally, as I have discussed, irrelevant features are never included in the semantic content of morphemes; null morphs are used for the purpose of describing non-overt realization of features; and "default" morphemes or "default" context specifications, in addition to a blocking principle, may be used in special circumstances creating a two-component grammatical structure: a lexicon that can overgeneralize and a filtering blocking principle that rules out overgeneralizations.

In the remainder of this chapter, I will begin considering a question of how a morphological lexicon of the sort discussed above might be learned. As a first stab at this question, I introduce an intuitive approach to learning form-meaning mappings. This approach, known as "cross-situational learning", has been informally discussed by many psychologists and linguists such as Pinker (1989); Fisher et al. (1994); Gleitman (1990) and others, and it underlies several computational models of word learning (e.g Siskind (1996); Thompson and Mooney (2003); Smith (2003); Vogt (2003)).

When applied to morphology, cross-situational learning runs into several problems. As I will discuss, these problems include null morphs, co-occurrence restrictions on morphemes, and homonymy. My main focus will be on tackling the last of these three problems - homonymy. Homonymy appears to be at first glance quite common in the domain of inflection, but as I show in chapter 4 the distribu-

tion of homonyms is not completely random - certain patterns appear to be more common than others. The formal learners I present at the end of this dissertation overcome the problem of homonymy and capture the statistical regularities in the data by predicting that those patterns that are rare are harder to learn.

2.2 Cross-situational approach to learning form-meaning mappings

In this section I introduce the general idea behind a basic cross-situational learner. The actual learner for learning form-meaning mappings of inflectional morphemes proposed in chapter 5 will be more complex, but it will build on the cross-situational strategy outlined here.

In the *Grundlagen der Arithmetic*, Gottlob Frege wrote "It is enough if the sentence as a whole has a meaning; it is this that confers on its parts also their content." This statement has been taken as a recipe for finding meanings of expressions (Hodges, 2000). The Fregean claim presupposes that languages have a compositional semantics and inspires the idea that one class of expressions is special because speakers have access to their meanings (e.g. sentences). The intuition is that meanings of "special" expressions can presumably be inferred from the environment (I use the term "environment" in its most general sense covering perceptual information about surroundings, inferences about speakers intentions, syntactic and distributional context of words, etc.).

A cross-situational approach to learning meanings is essentially a proposal about how to implement the Fregean idea, i.e., how to learn meanings of basic expressions from environments. For illustrative purposes I will introduce this approach in the context of learning word meanings, although soon after I will

switch to the problem addressed in this dissertation - learning of inflectional morphology. When applied to natural languages, cross-situational approach by itself is deficient for several reasons discussed here. But it will serve as a good starting point for understanding what properties of the input are particularly useful or problematic for learning.

I will take the *word* to be a "special" expression (whose meaning can be inferred from the environment) and the *morpheme* to be the basic unit of meaning. I also adopt a standard assumption that meanings of words can be usually derived compositionally from the meanings of their constituent morphemes.

2.2.1 Introduction to the cross-situational approach

Several constraints on the kinds of meanings human learners entertain as possible meanings of words have been proposed in the literature ("Whole Object Constraint", Markman (1989),"Mutual Exclusivity Constraint", Markman (1984)). However, while these constraints are certainly helpful, they are not sufficient for learning complex concepts. One needs further means for narrowing down the space of possibilities, especially since inferences drawn from only a couple of exposures to a word might be misleading.

One intuitive idea about how to narrow down potential meanings of a word involves keeping track of semantic properties that are constant across all contexts in which that word occurs. Imagine, for example, that a child is exposed to the word "car" when he is playing with his toy car, then when he sees a picture of a car in a book, and, finally, when he rides in a family sedan and sees other cars around him. The basic idea is that hearing the label "car" in all these different situations will help the child to abstract away from the irrelevant characteristics not included in the meaning of "car" (such as size, shape, color, model etc.) and

hone in on the more relevant characteristics such as "has four wheels," "has a steering wheel," "used to transport (toy) people and things," etc. (see however subsequent discussion of why this approach is not always appropriate especially for learning meanings of open-class items). This idea about how babies figure out what words mean is not a new one and is similar in spirit to the models of associative learning in which a connection between stimuli (experience) and a verbal response (words) is established and adjusted over time as associations between perceptual properties that always co-occur with the word strengthen, while other associations weaken (Skinner, 1957; Goldfarb, 1986; Regier, 2003).

Pinker (1989) suggests that verbs, just like nouns, can be learned through observations across different situations. He illustrates his point by considering verbs such as *fill* and *pour* that are used in very similar situations and whose meanings can be initially ambiguous for the learner. However, paying continual attention to the varying properties of the situations in which these verbs are used will help to disambiguate them. That is, the child will eventually experience the use of "pour" as opposed to "fill" in situations when the water is put in a glass up to the halfway point. On the other hand, the verb "fill" will eventually be used when a glass is left on the windowsill and is filled by the rain water. Based on such observations, the child will converge on the correct meanings.

Notice that the cross-situational approach to learning requires that the meanings of words can be exhaustively described in terms of some set of semantic primitives that combine to form more complex concepts (compositionality at the level of individual words). However, this notion of meaning is highly controversial. A more dominant view is that the meaning of a word (or, at least most words) cannot be defined in terms of a set of necessary and sufficient semantic primitives (Wittgenstein, 1953; Fodor et al., 1980; Fodor, 1998). Taking an

example from Wittgenstein, the word *game* is used in many different situations that taken together seem to have little in common (e.g., a chess game, a football game, a solitaire game, a game of wits and so on). According to Wittgenstein, if we look at all contexts in which the word *game* is used, we won't find any stable characteristics that pick out the class of games; instead we'll see "a complicated network of similarities overlapping and crisscrossing: sometimes overall similarities, sometimes similarities of detail." Fodor et al. (1980) present more general arguments against decompositional accounts of word meanings based on certain facts about reference fixing and informal inference. They also discuss a psycho-linguistics experiment that failed to show a relevant difference between causative verbs, thought to be semantically complex, and other "simple" verbs (although see a rebuttal of their arguments and critique of experimental design by Pitt (1999)).

However, when we look at the meanings of syntactic and grammatical complexes, such as sentences or sequences of inflectional morphemes, the situation is much less controversial. The meaning of a string of morphemes or a string of words is typically compositional. In fact, compositional accounts at this level correspond to the linguistic notion of grammar that, broadly speaking, specifies rules for combining structures and building larger expressions using finite means. This view is widely accepted as a way to understand the human ability to generate and comprehend the infinitely many grammatical expressions of a language. Thus, at these levels of grammatical structure, the compositionality requirement necessary for the cross-situational method is satisfied.

Decompositional analyses sometimes seem plausible even at the level of individual words or morphemes. For instance, such analyses have been proposed now and then for inflectional concepts like "person" and "number" which appear to

be complex, judging from their cross-linguistic realizations. I will discuss some such proposals in chapter 3 in connection to evaluating degree of homonymy and syncretism in the verbal agreement paradigms. A decompositional analysis is also appropriate and standard for morphemes that realize several inflectional features at once (cumulative exponence). For instance, the meaning of the verbal affix *-s* in English can be viewed as a complex "inflectional concept" that consists of a combination of several more primitive concepts such as "indicative," "present," "3 person," and "singular."

Another more practical concern raised in connection with the assumptions behind the cross-situational approach is the fact that in the real life situations, the immediate context to which a learner is attending does not always include relevant semantic properties that are denoted by the string. Bloom (2000) speculates that children are able to overcome this problem largely because they can often infer others' intentions by being particularly attuned to their gestures, facial expressions, intonation, following their eye gazes, and other types of information present in human interactions.[11] In addition, information from the neighboring words and syntactic context most likely also plays important role in aiding learning. So, we can take "situations" in the cross-situational picture of learning to mean something very general, covering variety of information sources mentioned above.[12]

We saw that cross-situational learning proceeds by keeping track of what prop-

[11] As discussed by Bloom, this hypothesis finds some support from the discrepant-looking paradigm experiments, where the experimenter utters a word while focusing her gaze on a different object than what the child is attending to (Baron-Cohen, et al. 1997). Normal and mentally handicapped children perform better at this task than autistic children who don't focus on human interactions. A certain percentage of autistic children are known to show a significant delay in vocabulary acquisition and other language skills.

[12] This assumption by itself is not entirely sufficient. There will be cases when a learner's inferences are incorrect or incomplete, and so the final learning algorithm would have to be robust enough to deal with noise. I leave the problem of noise to future research.

erties remain *invariant* across different situations and what properties change. If we think of situations in which words are uttered as sets of properties, then the invariant features of a particular morph can be found by taking *intersections* over all such sets. In the next section I discuss how this intersective strategy helps to zoom in on the meanings of individual morphemes by discarding irrelevant features. I also discuss how irrelevant features are connected to the notion of *underspecification*.

2.2.2 Irrelevant features and underspecification

The cross-situational learner not only must solve the mapping-problem (determining which morphs in a string correspond to which semantic features), but it also must identify which features present in the context are "extra" or irrelevant.

In general, irrelevant features are those features that have no effect on phonological realizations. For example, the feature "transitive" is irrelevant in English inflection because there are hardly any phonological contrasts due to different values of this feature, i.e., transitive and intransitive verbs are not inflectionally differentiated.[13] An alternative way of expressing complete irrelevance is to say that English does not inflect for transitivity.

Features can also be *partially irrelevant*, or irrelevant only in a certain context. For instance, animacy in Russian is partially irrelevant: most nouns in all cases and numbers don't inflect for animacy, except for declension 1 nouns. These nouns have different affixes in accusative depending on animacy. More specifically, for animate nouns, the accusative forms are identical to the genitive forms, while for the inanimate nouns they are identical to the nominative forms

[13] A few exceptions to this claim are verbs like "lay - lie" and "raise - rise". However, these verbs are often confused and used incorrectly by the native speakers which is a testimony to the unproductiveness of the transitivity as an inflectional category.

(the same can be said about agreement morphology on the adjectives).

Features that are irrelevant for the language as a whole result in uninflectedness, while those that are partially irrelevant result in *syncretism*, or in what I call *natural class syncretism* (see next chapter). Both uninflectedness and natural class syncretism are often described with *feature underspecification*. The way the term "underspecification" is used collapses an important for our purposes distinction, which I attempt to bring out by differentiating two kinds of underspecification: *strict* underspecification and *free* underspecification.

Strict underspecification rules out features that are either completely or partially irrelevant. A strictly underspecified feature matrix associated with some morph presents a set of necessary and sufficient feature values that describe the distribution of this morph. Notice that the above fact implies that affixes whose distribution can be described with strict underspecification are not homophones on my definition of homonymy. Another sign of strict underspecification is that the value of the underspecified feature never has an effect on the phonological realization in question. For example, animacy is underspecified for the English verbal third person agreement morpheme *-s*. That is, the 3rd person singular present tense verb will be marked with the suffix *-s* regardless of whether its subject is animate or inanimate. Strictly underspecified feature matrices correspond to partial functions of features to their values (or non-contradictory conjunctions of positive and negative literals), which are well understood mathematically and are known as *monomials*.

On the other hand, there is no correlation between irrelevance and another common use of the term "underspecification" that I call "free underspecification." Free underspecification is used in many morphological theories to describe morphemes that have an elsewhere-type distribution. For example, the present tense

verb-form *are* of the English verb *be* is often said to be fully underspecified for features of person and number. However, it is not true that either person or number are irrelevant features in the paradigm of the verb *be*. For instance, changing the value of number in the feature bundle [present, 1person, pl.] to [present, 1person, sg.] changes the phonological realization from *are* to *am*. Thus, the notions of strict and free underspecification are quite different. Free underspecification is usually used in tandem with blocking in order to correctly account for the distribution of the underspecified morphs. (Reasoning with blocking is a little trickier and less standard in formal theories than reasoning with monomials because it involves non-monotonicity.)

If for every form there is only one meaning in the language, then irrelevant features will be intersected out by the cross-situational learner. That is, in the absence of homonymy, cross-situational intersections are sufficient for solving the mapping problem and the problem of identifying irrelevant features (given that a few other properties hold in the input, see proofs in chapter 5). This fact by itself suggests a great advantage for languages with no homonymy - the existence of an extremely simple learning strategy for them. However, all languages contain instances of homonymy, which presents a problem for learning. In the next few sections I explores why homonymy is problematic, and briefly look at a few other phenomena that also provide a challenge for the cross-situational learner.

2.2.3 Homonymy as a problem for cross-situational learning

Recall that cross-situational approach involves making inferences over identical forms that occur across different situations. The first obvious case when this approach would fail is when identity of form does not imply identity of meaning and

instead is merely accidental, i.e. when the target language contains homonymy.

Reasoning across different situations based on forms that are homonymous will lead to overgeneralization (i.e., predicting a wider than actual distribution of morphs). For instance, consider a German verbal paradigm in 2.1 where the suffix *-e* occurs in several different verb-forms. It is evident from this example that the only feature value common to all contexts in which *-e* occurs is *singular*. Thus, a simple cross-situational learner would wrongly infer that the distribution of *-e* must be restricted to *singular*. This, however, would be an overgeneralization since there are many other singular contexts in which the morpheme *-e* does not occur.

Table 2.1: The present and past forms of the German verb "to play"

	present	past
1p.sg	spiel-**e**	spiel-t-**e**
2p.sg	spiel-st	spiel-te-st
3p.sg	spiel-t	spiel-t-**e**
1p.pl	spiel-en	spiel-t-en
2p.pl	spiel-t	spiel-te-t
3p.pl	spiel-en	spiel-t-en

Given that homonymy is problematic, we would like to know how common it is in natural language. For, if it is extremely common, the cross-situational approach is completely unfounded: as the German example above shows, the invariant features are not helpful in determining meanings of morphemes in the presence of homonymy. On the other hand, if homonymy is rather rare, then the cross-situational approach could still capture the majority of the data and something special could be done in the remaining cases.

It is clear that homonymy must be limited in some way. Imagine a language where every time you wanted to express a new meaning you would use exactly the same word. There would be no structure in such a language for the listener to

be able to infer anything about meanings of utterances. Nevertheless, it appears that homonymy is relatively common in inflectional paradigms,[14] although exact limits on it are generally not known. In this dissertation, I attempt to establish such limits by investigating how frequent homonymy occurs in verbal subject agreement paradigms.

As we will see in chapter 4, based on my calculations homonymy can be detected in roughly 25-30% of verbal agreement paradigms. Additionally, and more interestingly, I find that the cases of homonymy that do occur in inflection tend to be restricted in a particular way. This tendency has to do with the fact that homonymous patterns that cannot be described with blocking are particularly rare. The upshot of this statistical restriction is that it provides a clue for a learner about how to resolve most ambiguities arising due to homonymy.

To sum up, although homonymy is problematic for the cross-situational learner (it leads to overgeneralizations), we will see that the space of possibilities for attested form-meaning mappings is still structured in a way that makes learning easier. This idea is made more concrete in the chapter on learning.

[14]The fact that homonymy is relatively widespread in inflection might be related to a general tendency of homonyms to be prevalent among frequent lexical items. Ke (2004) analyzed the CELEX corpus for English, German, and Dutch and found that homonyms occurred in the highest frequency bands. He examined only a subset of homonyms, namely words that are spelled differently but pronounced the same. In English, 35 out of the 100 most frequent words were homophones, and 32 of them belonged to the closed class lexical items (cf. "I"/"eye," "to"/"too," "there"/"their," etc). Why would homonymy be more common in high-frequency lexical items? An intuitive explanation of his fact is that high frequency lexical elements are usually short, and hence allow for fewer phonological contrasts. Ke found some confirmation of this hypothesis by examining the degree of homophony among monosyllabic morphemes in 20 Chinese dialects. He found that dialects with the smaller syllable inventory had more homophony. Another possible reason for frequency of homonymy in inflection is the fact that phonological processes such as attrition, that lead to neutralization of phonological oppositions and eventually to homonymy, often occur at the word-edges where inflectional elements reside.

2.2.4 Other problems for cross-situational learning

Null morphs (or non-overt realization of meaning) and some co-occurrence restrictions on morphs also present an obstacle for the cross-situational learner.

Since it is impossible to directly observe when the null morphs occur in the string, it is impossible to rely on such occurrences for computing cross-situational intersections. Moreover, having multiple null morphs in the lexicon is like having homonymy that cannot be detected. And since homonymy is not handled by the simplistic cross-situational learner, neither are multiple null morphs. I discuss the problem of null morphs and possible solutions to it in chapter 5 section 5.2.3.

Depending on the particular mode of combination used for semantic values, some co-occurrence patterns among morphs could also be problematic for a learner that only calculates intersections. For example, if the meaning of the whole is exhaustively defined by the sum of the parts (a multiunion), then each semantic symbol that is part of the meaning of the string should be contributed by a single phonological symbol in that string. In accord with this assumption, given a word like *cran-berry* (provided that it consists of two parts), the meaning BERRY can only be attributed to one but not both of the parts. But the invariant values associated with *cran* will include BERRY since this morpheme always co-occurs with the stem *berry*. Setting aside the issue about what the appropriate analysis of *cranberry* is, if one is indeed aiming at representations where the same feature cannot be associated with more than one morph in a string, cross-situational learner by itself won't always be sufficient. It would have to be augmented with some additional inferences to further narrow down the hypotheses about morphs' meanings (see Siskind (1996); Kobele et al. (2003) for examples of such inferences).

My learner will avoid this issue all together because it will assume that the

appropriate mode of combination for inflectional affixes is unions rather than multi-unions (see more discussion on this point in chapter 5 section 5.2.1). Consequently, including the meaning BERRY as part of the meaning of the morpheme *cran* will not be problematic (but keep in mind that my algorithm applies to inflectional sequences rather than to the open class lexical items).

2.2.5 Synonymy and free variation

At first glance, the flip-side of homonymy, free variation or perfect synonymy does not appear to be problematic for the cross-situational learner.

Since we begin with distinct forms and proceed to generalize over the environments in which they occur, synonymy is handled straight-forwardly because it presents no ambiguity on the form-side. If we have two perfect synonyms, they will end up having exactly the same invariant features and, therefore, will be predicted to stand in free variation.

However, in a language with abundant homonymy, free variation can look very much like certain types of homonymy at an intermediate learning stage, especially when many of the irrelevant features have not yet been ruled out. This particular difficulty will be made more clear when we consider exactly what the learning algorithms in chapter 5 do in the presence of free variation.

In broad terms, the difficulty with free variation is that, at an intermediate learning state, it can be easily confused with a certain type of homonymy, that I call *overlapping* homonymy. The similarity between free variation and overlapping homonymy can also be seen from the fact that both are ruled out by the blocking proposals (for more discussion of this fact see section 5.7.3).

Although I will not concentrate as much on the problem of free variation, let me note here that it is a long standing idea that morphological doublets

are relatively rare in inflection, and when they do occur, they are historically unstable (Kroch 1994). Inflectional paradigms seem to be structured in such a way as not to allow more than one affix per paradigm cell. This is a much stronger restriction than a simple ban on perfect synonymy, as it also bars partial synonymy or taxonomic dependencies among affixes. That is, it is very unusual to see morphological systems where one could use one morpheme meaning "1 person, singular," or another morpheme meaning "singular" in exactly the same situation and in the same string (cf. words like "beans" and "legumes" in the open class vocabulary). Similarly, we don't normally see languages where some morpheme could be used to express "1 person" (any number) and another morpheme to express "singular" (any person) given that *person* and *number* are expressed cumulatively. If this were the case then we would expect that in the environments that included both "singular" and "1 person" feature values, either of the two morphemes could occur. (Note that in these examples the hypothetical morphs are not quite synonymous since at least one of them can occur in contexts where the other cannot.) Nevertheless, examples of free variation are attested and their learning has to be also eventually addressed.

CHAPTER 3

Constraints on form identity

At the end of the previous chapter I showed that homonymy presents a problem for learning morph-meaning mappings. A natural response to this problem is to observe that homonymy would not be a serious barrier for learning if it were restricted in a helpful way. For example, if homonyms always occurred in distinct contexts, then these contexts could be used to differentiate them.[1] The effects of context in inflectional morphology are briefly considered in section 3.1. However, most of the discussion in this chapter focuses on other kinds of restrictions on homonymy (and form identity in general) specified shortly.

The connection between the learner and the restrictions on homonymy is at least implicitly assumed in the literature on syncretism (Carstairs, 1984; Williams, 1994; Stump, 1993; Luraghi, 2000). In particular, this literature is concerned with the distinction between "systematic" and "accidental" identity of form, where "systematic" is most naturally interpreted as grounded in some principle that guides learning and language change.

Muller (2004) defines the notion of systematic as follows: "some instances of syncretism are ... systematic in the sense that they should follow from the morphological analysis." (p.197). The particular morphological analysis that Müller (and other proponents of DM) assumes draws a distinction between instances of

[1]For example, the cross-situational intersections could be taken within but not across different contextual domains.

syncretism that can be described with a single lexical entry (+ the assumption of blocking), and those that involve positing several lexical entries. It is the first kind of syncretism that is viewed as systematic. However, if the morphological analysis is completely disconnected from the analysis adopted by the speakers, then the statement that something is systematic just in case our theory says so, is meaningless. One coherent way to understand this statement is to assume that the morphological analysis is connected to the analysis imposed by the learners, so that whatever patterns can be easily captured by a particular morphological theory (e.g., because it involves positing a single lexical entry vs. several homonymous lexical entries) are also easily learned by the speakers and hence show signs of systematicity (such as relative stability, productivity, frequency, or whatever is normally meant by "systematic").

The first non-contextual restriction on form identity that I discuss has to do with an observation that many instances of identical phonological realizations in paradigms are due to neutralizations of partially irrelevant features (features that are underspecified in some context). Some authors want to restrict the term *syncretism* to this type of form identity only (Meiser, 1993). Syncretism defined in this way does not present an instance of true homonymy (as discussed in section 2.2.2). In the next chapter, I consider how common this type of syncretism is. This question will also help us determine a more general bound on homonymy in paradigms.

The second major restriction I consider explores an idea that seems to be particularly dominant in the Distributed Morphology tradition, namely that, among cases of inflectional homonymy, those that can be accounted for with help of blocking (along with a few other theoretical tools) are systematic (and hence common), while all others are accidental. This statistical restriction will also be

evaluated in the next chapter.

3.1 The effects of context

When adult speakers are faced with ambiguous input, the first intuitive idea about how they manage to resolve the ambiguity has to do with contextual clues that restrict the range of possible hypotheses, hopefully to a single most probable option. For example, when a geometry teacher speaks of an "angle dividing a plane," most students are not going to think of an airplane. The real world context (such as being in a geometry class), as well as the linguistic context (proximity to the words *angle* and *dividing*) will most likely be sufficient for disambiguating the word *plane*.

For the purpose of learning, if homonyms were indeed always restricted to contextually distinct domains in the adult language, this would potentially be of big help: once children learned to differentiate different contextual domains, they would be home free in terms of dealing with homonymy.

For example, if homonymy were possible across but not within different parts of speech, then by the time children learned to classify words into different parts of speech, they would be able to easily differentiate homophones. In fact, this is a standard technique used in natural language processing for word sense disambiguation for words like *paint* (noun) and *to paint* (verb). With development of part-of-speech tagging methods (whose state of the art performance is at 95%), homonymy that can be easily disambiguated by part of speech is no longer considered to be problematic. Thus, most word sense disambiguation models only focus on homonymy within the same grammatical category (Ide and Veronis, 1998). (Parts of speech are often determined based on syntactic information con-

tained in the neighboring words.) The fact that such a diambiguation strategy is successful makes it plausible as a technique that is at least on some occasions is also used by human comprehenders and learners.[2]

Another kind of obvious contextual difference that learners can potentially take advantage of is the syntactic position of morphemes (or words) within a larger phrase. Since in inflectional morphology the position of morphs within a word is largely linearly fixed, morph order provides an easy, string-evident, and tangible clue for differentiating homonymous affixes in different positions.[3]. For example, consider the following phrase in Aymara, an indigenous language of South America (from Hardman, 2001).

[2]To take advantage of this method, human learners would have to be able to differentiate parts of speech before they attempt to learn meanings, that is, in the absense of any semantic information. We don't know yet whether and how this can be done, but there are some proposals in the literature about first possible steps a learner can take to achieve this goal. In particular, Finch and Chater (1992); Mintz (2002) show that classifying words based on their occurrence in the same frequently encountered frames (such as [*was ... ing*] or [*the ... is*], etc.) correlates well with membership in the same grammatical categories. In reality, it could be that the two types of knowledge, grammatical category membership and form-meaning mappings, are acquired side by side in a bootstrapping fashion - knowing a little bit about grammatical categories can help to zero in on affix meanings, and vice versa, knowing the meaning of the affix might help to determine the grammatical category.

[3]However, there are some situations, where the position of an affix cannot be easily determined from the string because the string also contains some null morphemes. For instance, this could happen if all the slots in between the homonymous affixes are left empty, or if all the slots including the slot where one of the homonyms occurs are left empty. For example, it is not immediately clear whether the morph *-s* in the English string *cat-s* is the second-slot morph marking plurality or the third-slot morph marking possession (assuming that these two meanings are marked in different slots).

(1) kawk-sa-ru-sa

where-loc.side-direction-wh

"To which side"

In this language a suffix *sa* that occurs directly after the stem (usually attached to deictics and interrogatives) is a locational suffix meaning something like "side." A distinct but homonymous suffix *sa* marks information question either directly on the *wh* element or on the head of the *wh* containing phrase. Given their distinct distribution, the two *sa*'s are not likely to be confused with each other, even when they don't occur together in the same sentence (unless the situation discussed in footnote 3 arises).

The idea that distinct order of elements can resolve ambiguity is also familiar from the commonly observed tendency of languages with an impoverished inflectional system to develop fixed word-order. This is because when arguments are not distinctly marked, word order often (although not always) helps to determine their identity.[4]

Contextual disambiguation based on order within a string is easy to implement within the simple model of learning I propose in chapter 5. My learner will take the position of morphs within words into account, which automatically provides a way for disambiguating homophones occurring in different word slots.

Let me finally mention another type of distributional data that is potentially

[4]Even in languages with relatively free word order, certain orders become preferred in the presence of looming ambiguity. For instance, in Russian, as in many Indo-European languages, nominative and accusative inflectional markers are identical in some declensions. Word order is generally free in Russian, but sentences in which homonymy creates ambiguity between subject and direct object, are normally interpreted to have a fixed SVO order (Plank, 1980). For example,

Mat' l'ubit doč.

mother (nom/acc) loves daughter (nom/acc).

"The mother loves the daughter."

helpful in homonymy disambiguation. This data has to do with co-occurrence of ambiguous morphemes with other morphemes or free forms that mark some or all of the same distinctions unambiguously. To see how this might happen with bound morphemes, consider the data below from the New Guinean language Daga (based on the grammar by Murane, 1974). In this language suffixes occurring in several slots can have different shape depending on person and number. For instance, suffixes that occur in the second slot after the extended stem mark tense, person, and number. These suffixes can sometimes be followed by the so-called "medial suffixes" that make a medial verb (as opposed to a final verb). The paradigm of the verbal suffixes in the past tense for the conjugation A is given below:

Table 3.1: Daga past tense, class A suffixes (Murane, 1974)

	singular	plural
1	-an	-aton
2	-aan	-ayan
3	-en	-an

Notice that in the above paradigm 1st person singular suffix is homonymous with the 3rd person plural suffix. As mentioned before, the agreement suffixes can be followed by the medial suffixes which also have different allomorphs depending on person and number (see table 3.2).

Table 3.2: Daga past tense, medial suffixes (Murane, 1974)

	singular	plural
1	-a	-i
2	-a	-a
3	-i	-e

The medial suffixes are highly ambiguous, showing no distinctions between 1st person singular, 2nd person singular and 2nd person plural, as well as no

distinction between 3rd person plural and 1st person plural. However, when these suffixes are combined with the agreement suffixes, each person and number combination is actually uniquely determined (see table 3.3.)

Table 3.3: Past tense of the Daga verb *war* "to get"

	singular	plural
1	war-an-a	war-aton-i
2	war-aan-a	war-ayan-a
3	war-en-i	war-an-e

The learner that keeps track of the co-occurrence patterns among morphs could take advantage of them in disambiguating the types of homonyms discussed above. For example, such a learner could adopt a conservative generalization strategy, such that it will initially treat morphs that have different co-occurrence patterns as belonging to different sub-classes. This learner will not generalize across such sub-classes, until it determines that it is safe to collapse them.[5]

It remains to be quantitatively shown that inflectional homonymy in general is more prevalent when it can be contextually disambiguated.[6] But even if it were the case, this would still not make such homonymy trivial to learn. This is true in particular because at the early stages of learning children might not have a strong sense of what constitutes a different contextual domain, or they might not be able to integrate many sources of information in trying to interpret and decipher meanings of words. Some psycholinguistic studies suggest this might

[5]My learner will not be as sophisticated, it will only keep track of the morphs' positions and not of their co-occurrence patterns.

[6]Plank (1980) shows that this might at least be true for particular constructions. He studied possessive constructions in which there was a danger of an identical encoding of the possessee and the possessor. There seems to be a strong tendency in a number of languages he considered for keeping this contrast distinct while allowing neutralizations of semantic contrasts in many other areas, including the subject-object contrast encoded by case. For example, in languages such as Finnish and Uzbek, the word order in possesive constructions is generally free, except in cases where the genitive is syncretic with the nominative. In such cases, the strict word-order presumably indicates the default interpretation of the possessor possessee relationship.

be true for some non-distributional, pragmatic notions of context. These studies show that children do not process information in the same way as adults do, and in many cases fail to rely on pragmatic information in parsing ambiguous input (Trueswell et al., 1993).

Additionally, most cases of inflectional form identity that receive a great deal of attention in the literature come from affixes that occur in the same word slot, and that have similar distributions, i.e., affixes that usually belong to the same sub-paradigm. Such cases of homonymy/syncretism are less likely to be easily accounted for by contextual differences. Therefore, we still need an alternative strategy of dealing with homonymy. The next two sections begin to investigate restrictions that will eventually help us to formulate such an alternative strategy.

3.2 Natural class syncretism

When we look at inflectional paradigms as they are traditionally represented in grammars, we notice that often the same phonological form will occupy more than one paradigmatic cell. Such inflectional identity has been the subject of many papers seeking to define what instances of identity should count as systematic and what instances should count as truly accidental (the use of these terms differs depending on the theory). Systematic form identity is generally referred to by the term syncretism. In the diachronic perspective, syncretism is used to refer to the process of neutralization of some semantic (or even syntactic) contrast(s) which results in the phonological merger of several inflectional markers (Bazell, 1960; Luraghi, 1987). In other words, conceptually similar morphological categories (e.g. "plural" and "dual") might over time be re-analyzed as a single category and hence be expressed by the same formal means. The idea that syncretism reflects semantic relatedness has been present in the literature for a long time.

It is at least implicitly assumed in many proposals about semantic organization of inflectional features, for instance in Jakobson's famous analysis of the case semantics informed largely by the syncretism in the Russian nominal paradigm (Jakobson, 1936).

If a semantic contrast is completely lost overtime, it would not occur to us to describe the resulting pattern of identity as homonymy. Complete loss of contrasts is simply interpreted as irrelevance or absence of some semantic distinction. But if a contrast is neutralized only in a particular sub-paradigm of the grammar, this looks more like homonymy since the same means are used to express several inflectional concepts that are differentiated elsewhere in the language. Yet, in essence, there is no deep difference between partial and full neutralization of contrasts. Both can be described with underspecification, neither is an instance of homonymy, and hence neither is problematic for the cross-situational learner described earlier. For instance, if categories of gender such as "masculine" and "feminine" are merged in the plural, but remain distinct in the singular, the gender features will be intersected out in the plural contexts only.

Partial neutralization of an inflectional contrast could be linked to a historical re-analysis in which semantically similar morphological concepts were merged in particular sub-paradigms (or in the presence of some other features). From the synchronic point of view, however, speakers are not aware of which instances of non-distinction came about via a historical process of semantic neutralization and which were mere accidents. What matters from their point of view is that some instances of form identity *look* as though they arose via a systematic merger (i.e., the syncretic categories form a semantically natural class definable by a necessary and sufficient set of feature values), while other instances of identity do not. In the first case, the syncretic morph has a homogeneous and systematic pattern of

distribution and can be said to have a single meaning. I will refer to this type of identity as *natural class* syncretism.

If most cases of paradigmatic form identity were due to natural class syncretism, this would mean that inflectional paradigms tend to avoid homonymy (this is stated in the hypothesis below). The next chapter will address the question of whether this hypothesis is actually true.

Hypothesis 1: Natural class syncretism, as well as full non-distinction of contrasts should be historically stable and relatively common cross-linguistically compared to all other types of form-meaning mappings in inflectional paradigms (i.e., mappings involving homonymy).

3.3 The elsewhere and the overlapping homonymy

In the previous section, we saw that some instances of form identity are due to natural class syncretism. We have put forth a hypothesis that perhaps most cases of indenty are of this type. In this section, we will consider a possible statistical restriction on the remaining types of form ambiguity. Namely, we formulate a hypothesis that among the homonymous mappings the cases that can be described with defaults are particularly common. This proposal is at least implicitly present in many morphological theories that make use of blocking (Distributed Morphology, Paradigm Function Morphology, Network Morphology and others). In such theories, affixes that can be described by a single lexical entry (or a singe rule) that has an "elsewhere" status with respect to some other set of representations or rules are are believed to present instances of systematic (rather than accidental) homonymy.

The idea of defaults and "elsewhere" patterns has played a crucial role in grammatical descriptions since Panini. It provides an intuitive short-hand for capturing what appears to be the set of "left-over" items that do not fit into any other well-defined category. Below, I discuss how the notion of defaults is instantiated in the framework of Distributed Morphology, and how it ties into the distinction between accidental vs. systematic homonymy.[7]

In Distributed Morphology, defaults are achieved by means of free underspecification, the Subset Principle, and sometimes the so-called Rules of Impoverishment. To quickly demonstrate how these mechanisms work, consider the following example (from the Distributed Morphology website, Sauerland (1995)).

Table 3.4: Distribution of adjectival suffixes in Norwegian, Sauerland (1995)

STRONG (used with definites)	-neuter	+neuter
-pl	zero	t
+pl	e	e
WEAK (used with indefinites)		
-pl	e	e
+pl	e	e

In Norwegian, the weak adjectival ending *-e* does not differentiate number or gender. Additionally, this ending is homonymous with the plural ending of strong adjectives. Sauerland proposes the following analysis for the Norwegian adjectival suffixes.

(2) DM-style lexical entries for Norwegian

zero – [-pl, -neut]/Adj +

t – [-pl, +neut]/Adj +

e – elsewhere/Adj +

[7]Notice that defaults do not have to be global, they can be relativized to a particular corner of a paradigm.

In this analysis, *-e* is underspecified and has the elsewhere distribution. It is used in the plural contexts because it is the only morpheme that is compatible with the specification [+pl] by virtue of being underspecified for number. The fact that this morpheme is used in the "weak" contexts is explained by appealing to Impoverishment. More specifically, when the adjective is used with an indefinite noun ("weak" syntactic position), a post-lexical rule of Impoverishment is said to delete the gender features from the syntactic representation. This deletion has an effect of blocking morphemes specified for gender from being inserted into the corresponding morpho-syntactic representation and triggering what Halle and Marantz call "retreat to the more general case," or the insertion of the default morpheme into the "impoverished" syntactic node (for more details on Rules of Impoverishment see Noyer (1998)).

Let me point out that it is possible for a single paradigm to contain several default or "elsewhere" morphemes. (Roughly speaking, a default morpheme is a morpheme whose distribution is described with a freely underspecified feature matrix and a Blocking Principle.) This can happen if the defaults are nested within each other, or if they share the same "blocker(s)", or if they are completely disjoint. These possibilities are demonstrated schematically in figure 3.1 in (1) (2) and (3), correspondingly. I use the notation "A $\gg$ B" to indicate that morpheme A blocks morpheme B. This means that B occurs in the box labeled B (representing some natural class of meanings) except where this box overlaps with the box labeled A.

In the case of the nested distribution in (1), the default morphemes are B and C, where B is an elsewhere case with respect to A, and C is an elsewhere case with respect to both B and A. In case (2), A and B are defaults with respect to the same morpheme C. Finally, in case (3), there are two disjoint defaults C

(with respect to D) and A (with respect to B).

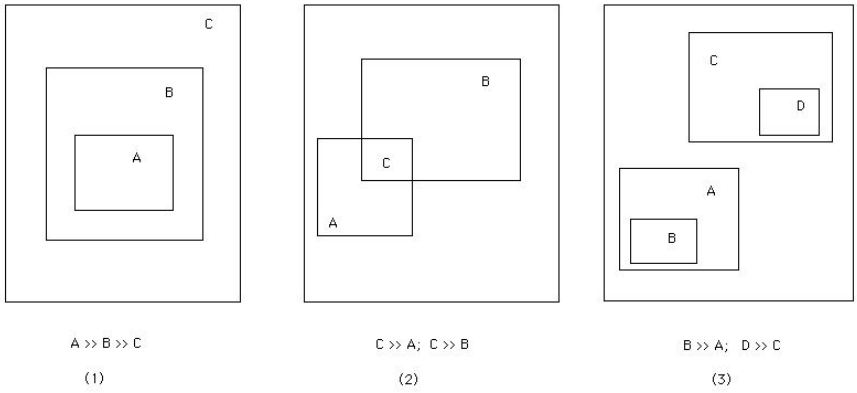

Figure 3.1: Cases of multiple defaults within a single paradigm

Also note that Rules of Impoverishment (which were proposed as an alternative to rules of referral (Zwicky, 1985; Stump, 1993) are extremely powerful. The only restrictive power they have comes from a stipulation that marked features (and features that depend on them) are more likely to be "impoverished" (deleted from the syntactic representations) than unmarked features. (However, this connection to markedness is not specific to the impoverishment mechanism per se and could be build into any other theory, including the rules of referral). Since Rules of Impoverishment are not very restrictive, they don't provide a good way of constraining homonymy.

On the other hand, the Subset Principle and blocking proposals in general are in principle restrictive since not all patterns of homonymy can be described by appealing to defaults. In particular, certain patterns that I call *overlapping* are not amenable to an analysis in which every morph is assigned a single lexical value and some morphs have a default status. Figure 3.2 schematically depicts such overlapping patterns. They can be either due to overlapping homonymy or to

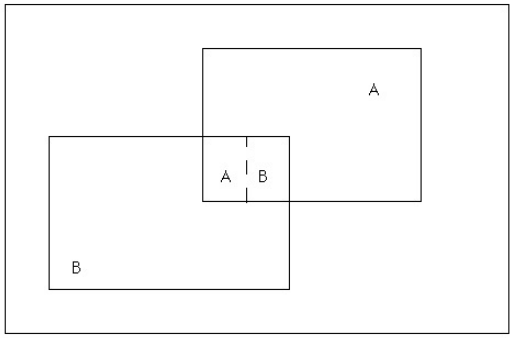

Figure 3.2: Overlapping Homonymy

free variation. More concretely, this picture shows a situation in which A and B occur in the same sub-paradigm that lies at the intersection of the natural classes that limit the distribution of A and B (see a more precise definition below).

To define overlapping distributions I first need to define a notion of *invariant features*, and for that I need to make the terms *paradigm* and *paradigm cell* more precise.

Definition of the Overlapping Affix Distribution

1. A *paradigm* over a set of features F is a collection of all complete assignments of features in F to their values.

2. Each complete assignment is called a *cell*. For example, if F includes two features *gender* (with values "masculine" and "feminine") and *number* (with values "singular" and "plural"), then the combination [masculine, singular] is a complete assignment and one of the four cells in a paradigm over F.

3. When we say that a morph is associated with a cell, this means it expresses features of that cell.

4. If we take an intersection of all cells occupied by some morph m, we will get some set of feature values that I call *invariant features* of m or I(m).

We can now define exactly what it means for two morphs to stand in an overlapping distribution.

(3) Two morphs x and y, are in the *overlapping* distribution just in case the two conditions below are met:[8]

a. x and y are in competition. That is, the invariant features of x are *consistent* with the invariant features of y, which is to say that $I(x) \cup I(y)$ contains no contradictory features.

b. x occurs in the domain of the invariant features of y and vice versa. That is, $\exists$ a cell c in a paradigm, where $c \supseteq I(x) \cup I(y)$, such that c is associated with x, and $\exists$ a cell c in a paradigm, where $c \supseteq I(x) \cup I(y)$, such that c is associated with y.

For an example of an overlapping distribution, consider the German paradigm for regular verbs below (I assume *person* features such as "participant in the speech event" and "speaker", and a *number* feature "group").

Table 3.5: Present tense paradigm of the German regular verbs

	number	
person	sg: -group	pl: +group
1p: +part,+speak	-e/-ø	-en
2p: +part,-speak	-st	-t
3p: -part,-speak	-t	-en

In this paradigm, the invariant features of the affix *-en* (that is, the features that are present in *all* cells where this affix occurs) is the set [+group]. The

[8] As it will be made more explicit in Chapter 5, the overlapping relation is transitive so that if x and y are overlapping and y and z are overlapping, then x and z are also overlapping.

invariant feature of the affix *-t* is [-speaker]. These two feature sets are non-contradictory and therefore consistent with each other. In addition, *-t* occurs in the domain of the invariant features of *-en*, i.e. in one of the [+group] cells (2p.pl.), and vice versa *-en* occurs in the domain of the invariant features of *-t*, i.e., in the [-speaker] cell (3p.pl). Therefore, this is an example of overlapping homonymy. The above paradigm contains another overlapping distribution due to free variation in the 1p.sg. cell. Neither *-e* nor *-ø* are homonymous in the present paradigm: each can be described in terms of a single natural class of features. Nevertheless, their invariant features are consistent, and both occur in exactly the same domain - 1p.sg. (i.e. they meet the definition of the overlapping distribution).

The overlapping distributions are precisely those that cannot be described exclusively with free underspecification coupled with the Subset Principle. This is because neither of the overlapping morphs can be said to block the other morph since either both of them occur in exactly the same cell (free variation), or each blocks the other in some cell (overlapping homonymy).

To the extent that overlapping distributions (free variation and overlapping homonymy) are empirically attested, they are usually viewed as idiosyncratic or accidental. Such patterns are hypothesized to be historically unstable and hence in some sense non-optimal or difficult for the speakers to learn. The fact that such patterns actually exist has been pointed out before, but as far as I know, there has been no typological investigation of their relative frequency. I believe this is largely due to fact that there was no clear understanding of exactly what types of affix distributions cannot be described with blocking. Once we have a precise formulation of such distributions (provided above) we can ask the question of how frequent are overlapping patterns compared to the elsewhere homonymy

and natural class syncretism.

Since, intuitively, overlapping homonymy appears to be more accidental and more complex, the particular hypothesis that we'd like to evaluate next is that homonymy in inflectional paradigms is rarely due to overlaps and most often can be described as an "elsewhere" case. If this is indeed cross-linguistically true, this would be beneficial for a learner biased to use simpler learning strategies whenever possible.

Hypothesis 2: Elsewhere cases of homonymy are historically more stable and more common cross-linguistically compared to the overlapping homonymy which is expected to be rare in inflectional paradigms.

This hypothesis as well as the first hypothesis will be evaluated against typological data in the next chapter. In the next few subsections I give more examples of overlapping distributions of three different types. The reader can skip these sections without any loss in continuity.

3.3.0.1 Overlapping Homonymy Type 1: same invariant features

The first obvious case which allows to easily detect an overlapping homonymy has to do with paradigms where two (or more) homonyms have exactly the same invariant features (i.e., the intersections of all cells in which they occur are the same). An example of this case comes from the verbal paradigm of the Cushitic language Dhaasanac (Baerman (2004), described by Tosco 2001). All verbs in this language distinguish two stems, which are abstractly labeled as A and B in the table below (an example verb is given in parenthesis).

Notice that if we take the intersections of all fully specified cells occupied by the A form, it will be equivalent to the intersection of all cells occupied by the B

Table 3.6: Dhaasanac verbal paradigm, example verb: kufji - kuyyi "to die"

		sg	pl
1p	(incl)	–	A (kufi)
	(excl)	A (kufi)	B (kuyyi)
2p		B (kuyyi)	B (kuyyi)
3p	(masc)	A (kufi)	A (kufi)
	(fem)	B (kuyyi)	A (kufi)

forms, because both A and B occur with all possible person, number, and gender values. So the first condition for overlapping homonymy in the definition (3) is obviously met. It is not hard to see that the second condition, the overlapping property, is met as well. We can verify that the Dhaasanac distribution of stem allomorphs cannot be described by appealing to defaults, since it is impossible for both A and B to be defaults with respect to the same sub-paradigm.

3.3.0.2 Overlapping Homonymy Type 2: equally specific invariant features

Another example of the overlapping homonymy comes from the subset of paradigms in which homonymous affixes have different invariant features that are consistent with each other and neither of them is more specific than the other.[9] The German paradigm in table 3.5 presents one example of this homonymy. For another example consider the verbal paradigm of French conjugation I verbs in the future

[9]Not all cases of this sort involve overlapping homonymy. Some paradigms in which invariant features of several homonymous morphs are consistent and equally specific are of the "elsewhere" type. Because they are not covered by the Subset Principle the way it is defined, there have been proposals in the literature in which the competition between equally specific vocabulary items is resolved either by language particular preferences (see Halle and Marantz, 1993; Hjelmslev, 1935) or alternatively by a universal feature hierarchy (UFH, Noyer, 1998). The vocabulary item that is specified for a feature appearing higher in the UFH or that is stipulated to be more important by a language specific preference rule, wins the competition. Yet, as this section demonstrates, neither language specific nor the universal hierarchy of features can account for all patterns of homonymy in which the invariant features of two competing morphemes are equally specific.

tense (given in phonetic transcription in table 3.7).

Table 3.7: French, conj.I. future tense suffixes

	number	
person	sg: -group	pl: +group
1p: +part,+speak	-Re	-Rõ
2p: +part,-speak	-Ra	-Re
3p: -part,-speak	-Ra	-Rõ

Observe that the suffix -*Rõ* always occurs in the cells whose intersection yields [+group], and the the suffix -*Re* occurs in the cells whose intersection is [+participant]. These two feature sets are consistent with each other and neither is more specific than the other. Moreover, they meet the overlapping condition: [+group] is consistent with a cell occupied by -*Re*, (i.e. [+part., -speaker, +group]), and [+participant] is consistent with a cell occupied by -*Rõ* (i.e. [+part, +speaker, +group]). No feature hierarchy will help us here: we cannot stipulate that [-sg] should override [+participant] or vice versa because blocking is happening in both directions. One could perhaps reanalyze this example using different feature values, but the point remains - no blocking principle (even if it involves feature hierarchies or stipulated blocking relationships) can in principle resolve all conflicts between equally specific items.

3.3.0.3 Overlapping Homonymy Type 3: invariant features in the subset relation but in the wrong direction

Finally, the last example of overlapping homonymy shows that even when invariant values of competing morphemes *stand in a subset relationship to each other*, they cannot always be accounted for by the Subset Principle. This case is illustrated in the sub-paradigm of the Slovenian pronominal adjective "that" in table 3.8 that encompasses non-oblique cases only (nominative and accusative).

Table 3.8: Slovenian pronominal adjective "that"

	sg			du			pl		
	masc	neut	fem	masc	neut	fem	masc	neut	fem
nom.	tâ	tǫ̂	tâ	tâ	tę̂	tę̂	tî	tâ	tę̂
acc.	tâ	tǫ̂	tǫ̂	tâ	tę̂	tę̂	tę̂	tâ	te

Calculating the invariant features for all different forms of "that", we get the following.

(4) The invariant features for forms of "that"

tî [-oblique case, +nom, -sg, -du, +masc]

tǫ̂ [-oblique case, +sg, -du, -masc]

tę̂ [-oblique case, -sg]

tâ [-oblique case]

Given the above paradigm and the list of invariant feature values, the reader can verify that the forms *tę̂* and *tâ* meet the requirements for the overlapping homonymy. Correspondingly, Slovenian data cannot be explained by the Subset principle. In fact, the Subset Principle applied to the lexical representations based on the invariant features would make a wrong prediction in this case. More specifically, it would predict that *tę̂* should appear in the contexts [-oblique case,-sg,+du,+masc] because it is the most specific affix compatible with this context. But in fact, the form that actually appears there is *tâ*. (Similarly, *tâ* also overlaps with *tô*.)

One important conclusion we can draw from this discussion is that a simple learning algorithm that finds invariant features by underspecifying whenever possible, and that applies the Subset Principle to resolve the resulting conflicts, will not do. It will fail to correctly account for instances of overlapping homonymy.

The question of what the learner should do in such cases will be taken up in chapter 4.

CHAPTER 4

Evaluating constraints on form identity

In this chapter, I take a look at whether the restrictions on form identity discussed in chapter 3 are supported by typological data. Given that these restrictions are hypothesised to be connected to the learning algorithm, the rationale behind considering typological frequencies rests on the common assumption that patterns grounded in learning biases should be more frequent cross-linguistically than the merely accidental patterns. The intuition behind this idea is this: presumably the learner will have more difficulty in learning accidental patterns and over time such patterns should be regularized and appear less frequently.[1]

Recall that there are two claims that we would like to evaluate in connection to phonological realizations of semantic contrasts. These claims are repeated below.

(1) Hypotheses for evaluation

a. Form identity due to homonymy is relatively rare in morphological paradigms. That is, most mappings between form and meaning are either unambiguous or are instances of natural class syncretism (due to partial neutralization of some semantic contrasts).

[1]In reality, things are a bit more complicated. For instance, even if accidental homonymy diminishes over time, new instances of it emerge afresh. More generally, there are other factors besides learning biases that affect language change and the amount of affixal ambiguity. Nevertheless, it could still be that a particular learning mechanism is biasing the empirical data which is reflected in strong typological preferences, if not in categorical restrictions.

b. Among attested instances of inflectional homonymy (i) elsewhere patterns are common, and (ii) overlapping patterns are rare

Typological studies investigating constraints on inflectional identity are not numerous, but they do exist. The most comprehensive of such studies is perhaps presented in the recent book by Baerman et al. (2005), based on typological work of the Surrey Morphology group. This book attempts to evaluate various notions of syncretism proposed over the years against a typologically diverse sample of languages (based in part on the sample from the World Atlas of Language Structures and in part on the authors' own databases). The data used in this sample has been compiled into the Syncretism Database available on-line (at www.smg.surrey.ac.uk/Syncretism/index.aspx).

In their book, Baerman et al. briefly discuss the proposal that syncretism reflects an underlying organization of semantic features into natural classes. However, the models of feature structure they discuss include the "elsewhere" syncretism into the category of natural class, unlike the terminology I have adopted. In these models, as long as syncretism can be described with feature underspecification (whether "free" or "strict") it is considered to reflect a natural class grouping (the blocking relationships are implicitly assumed). Baerman et al. give several examples that cannot be accounted for in terms of underspecification on any assumption about the feature structure[2], although they don't address the question of exactly how wide-spread such patterns are.

In this chapter, I will take a closer look at this question, as well as the question of the overall frequency of homonymy in inflection. In particular, I will

[2]The only exception is the so-called *cross-classifying* model of features from Johnston 1997. However, from what I can understand, this model allows one to group any features whatsoever into a "natural class" and thus is completely nonrestrictive.

examine verbal agreement paradigms of 30 genetically and geographically diverse languages. This sample of languages was largely based on the Surrey Person Syncretism Database mentioned above. Before I discuss the typological data, I lay out a feature analysis of agreement morphology that will serve as a frame of reference for evaluating semantic relatedness of inflectional contrasts.

Also, before I turn to the typological data, I attempt to address a frequently neglected issue of the expected chance frequencies of different affix distributions. Evaluation of actual frequencies should take the chance frequencies into account to make sure that some pattern is not frequent for a trivial reason, namely that it is expected to be frequent by pure chance. Calculating chance frequencies is challenging for several reasons, the most significant one being that we don't quite know what the appropriate underlying feature structure of universal inflectional contrasts looks like. Thus, the calculations I present here serve only as very crude approximations of the actual chance frequencies.

4.1 Computing chance frequencies

Given a paradigm of a particular size defined by a particular feature system, certain types of homonymy will be more likely than others for purely combinatorial reasons. To give a very simple example, if we have a two cell paradigm defined over a single binary feature, there are only two possible ways to fill it, and none of them could constitute an overlapping homonymy. This is because an overlapping homonymy has to involve at least two ambiguous morphemes, and this is impossible in a paradigm with two cells.

In order to rule out the possibility that some patterns of form identity are cross-linguistically frequent or rare for combinatorial reasons, we would like to

have at least an approximate estimate of the expected or "chance" frequencies of these patterns as a function of the paradigm size. We can conclude that a particular type of mapping is favored if its actual frequency is much higher than its expected frequency.

To calculate the exact expected frequencies, we would need to know what the right feature system is, what (if any) dependencies among the features there are, and what the maximum number of affixes a paradigm may have. Unfortunately, we don't have a very good theory of these facts (as I will discuss in the next subsection). So, as a first approximation, I will calculate the expected frequencies for systems defined over n binary features with an assumption that all features are independent. These calculations present an upper bound for the types of feature systems most commonly assumed in linguistics – that is, systems including dependencies among features (e.g., the feature "tense" is dependent on the feature value "+ finite" meaning that it can only be activated in [+finite] environments). When a feature system includes dependencies of the sort above, it can be described as a paradigm with "gaps" where certain combinations of features are impossible.[3] So, a paradigm defined over a feature system with dependencies will have less cells than a paradigm defined over the same number of features which are fully independent. As a result, the assumption of independence implies that we will be computing upper bounds for the expected frequencies of different affix distributions. I will comment on this fact later when we consider the specific results for the chance frequencies.

Additionally, for simplicity I assume that every cell may be occupied by at most one affix (i.e., free variation is ruled out). Note that with no bound on free variation, the number of possible affixes and possible affix arrangements is

[3]By a "gap" here I mean a logically impossible combination of features, rather than an accidental absence of phonological realization for some logically possible distinction.

infinite in the worst case scenario.

So, our null hypothesis is that given a paradigm with n cells and an inventory of n affixes, any distribution of affixes such that each paradigm cell is filled by some affix, but not every affix is necessarily assigned to a cell, is equiprobable. The first question we want to ask is this: what is the total number of possibilities for assigning affixes to paradigm cells? This question is answered in the next section.

4.1.1 Total number of possible mappings

Ultimately we are interested in finding out the expected frequencies of paradigms that contain instances of homonymy vs. those that don't, and the expected frequencies of paradigms with overlaps vs. those without. To calculate these expectations, we would first need to know the total number of possible mappings between affixes and paradigm cells under the null assumption that any mapping is equally likely.

In the absence of free variation, we can view a paradigm arrangement as a partition induced by affixes, where the cells occupied by the same affix are grouped into the same partition block. The maximum number of affixes in a paradigm with n cells is n, and the minimum number of affixes is 1. This view helps us see that the number of all possible arrangements of affixes in a paradigm of size n is equivalent to the number of partitions (or the number of equivalence classes) of an n size set.[4] This number equals to the sum of the "Stirling numbers

[4]Notice that this view gives us a way of calculating possible paradigm types, ignoring identity of actual affixes. That is, paradigms that are isomorphic to each other up to relabeling of the affixes are counted as the same. An alternative way of calculating the space of possibilities would be to consider all possible mappings for a particular inventory of inflectional affixes. For example, if we could define a finite universal set of possible affixes using a universal set of phonemes, we could ask how many ways there are of mapping this particular set to a particular number of cells such that every cell has only one affix in it, but not every affix need be mapped

of the second kind," and is also known as the Bell number (Rota, 1964). The Stirling numbers (abbreviated S(n,k)) give us a number of partitions of a set of size n into k subsets. The formulas for the Stirling and the Bell numbers are given below.

(2) Stirling numbers and the Bell number

$$S(n,k) = \frac{1}{k!}\sum_{j=1}^{k}(-1)^{k-j}\binom{k}{j}j^n$$

$$B_n = \sum_{k=1}^{n} S(n,k)$$

For example, the number of all possible arrangements of affixes in a paradigm with 3 cells is $B_3 = 5$. I list all these arrangements in figure 4.1.

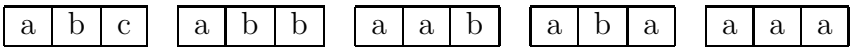

Figure 4.1: Partitions of size 3

Bell numbers grow extremely fast (consider their values for the first few n's in table 4.1).

Table 4.1: Bell numbers

number of cells, n	number of different partitions, B(n)
3	5
4	15
5	52
6	203
7	877
8	4,140
9	21,147

These numbers alone are already suggestive. If any pattern of affix assignment were equally likely, then even for relatively moderate-sized paradigms, we

to a cell. This would be similar to calculating the number of all possible paradigm tokens.

would expect to see a huge number of different paradigm types with no particular consistencies in realization of paradigmatic contrasts. However, although all languages contain inconsistencies and ambiguities, even an untrained eye can notice that majority of the contrasts in languages are made along the lines of the natural classes formed by the features. For example, from language to language we see that meanings such as "singular" are realized consistently by the same morph or a set of morphs which are phonologically distinct from the realizations of the meaning "plural." This observation is essentially at core of the first hypothesis in (1), which is considered in greater detail in the following section.

4.1.2 Expected occurrence of paradigms with no homonymy

In this section we estimate the expected proportion of paradigms that contain no instances of homonymy. Recall that non-homonymous mappings are those in which the distribution of each affix can be precisely defined in terms of a single set of necessary and sufficient feature values, comprising a natural class of meanings associated with the same affix. Thus, to calculate the number of paradigms with no homonymy, we need to calculate partitions in which each block forms a complete natural class.

Recall that we assume features to be binary and independent. The number of natural classes for a paradigm defined over n binary independent features is equivalent to the number of ways in which a feature matrix can be underspecified. This number is 3^n given that each feature could range over three possible values: $+$, $-$ or underspecified. There is however no mathematical formula (that I am aware of) for finding the number of partitions in which each block forms a natural class. To calculate this number, I used a program written by Jeff Heinz for the

purpose of generating natural class partitions of phonological features.[5] The results of these calculations are shown in table 4.2.

Table 4.2: Expected proportion of paradigms with no homonymy

feat. n	cells 2^n	# of nat. classes 3^n	# of parad. with no homonymy - (no formula)	# of possible partitions B_{2^n}	p. of parad. with no homonymy
1	2	3	2	2	1
2	4	9	8	15	.53
3	8	27	146	4,140	.03
4	16	81	61,712	$104{,}8 * 10^5$	$5.8 * 10^{-6}$

As you can see, the proportion of paradigms with no homonymy (i.e., paradigms in which each partition forms a natural class) decreases very rapidly as the size of the paradigms grows. For instance, in a paradigm with 8 cells defined by 3 binary features, the percentage of affix arrangements that contain no homonymy is already less than 5 percent. In other words, if affixes were distributed in a completely random way, then paradigms with no homonymy would be very rare. Recall that these calculations present an upper bound for systems containing dependent features. Thus, if morphological systems include feature dependencies, the expected number of paradigms with no homonymy is even lower than the estimations above suggest.

This result is perhaps not so surprising: we would not expect languages to respect the 1-1 correspondence between forms and meanings by pure chance. It seems only natural that the 1-1 property of semantic mappings is systematic and would be difficult to obtain randomly.

[5]The software that I used for this calculation can be obtained from http://www.linguistics.ucla.edu/people/grads/jheinz/software/index.html).

4.1.3 Expected occurrence of paradigms with overlapping and elsewhere homonymy

In the previous section, we saw that fully unambiguous paradigms are not expected to be frequent by chance. Thus, an overwhelming majority of affix arrangements occurring completely randomly are very likely to involve homonymy. The question we ask in this section is how many of these arrangements are expected to include at least one instance of overlapping homonymy? Those that don't include any overlapping homonymy must be cases of elsewhere distributions (describable by blocking and underspecification).

Once again, there is no easy formula for calculating the proportion of arrangements including overlaps. I did it by writing a computer program that first generated all possible partitions for a set of cells defined by n binary features, and then counted how many of these partitions included cases of overlaps. As before, to calculate the expected frequency, we will divide this count by the total number of possible arrangements given by the Bell number. The proportion of the elsewhere arrangements is found by subtracting the expected proportions of paradigms with overlapping homonymy and the expected proportions of paradigms with no homonymy from 1. The results of these calculations are summarized below.

Table 4.3: Upper bounds on overlapping homonymy

feat.	cells	p. of parad. with no homonymy	p. of parad. with overlaps	p. of elsewhere paradigms
1	2	1.	0	0
2	4	0.53	1/15=.06	.41
3	8	0.03	2,658/4,140=.64	.33

We see that the proportion of paradigms containing at least one instance of overlapping homonymy grows extremely fast (in reverse proportion to paradigms

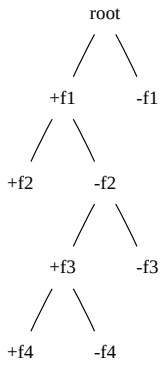

Figure 4.2: A feature hierarchy with dependencies

with no homonymy). Thus, the bigger the paradigm, the more we can expect it to include an instance of overlapping homonymy by chance.

However, these estimations are not particularly telling if the actual frequency of overlapping distributions turns out to be rather low. This is due to the general fact that upper bound estimates are only useful if the observed frequencies exceed them. The worry is that if overlapping patterns are empirically rare, it could still be the case that they are expected to be rare given feature systems with many dependencies. To check this hypothesis, I tried to estimate expected proportions of overlapping homonymy in such systems. To take an extreme case, I looked at the feature systems with the maximum number of dependencies, i.e., systems in which all but one feature are dependent. Such feature systems look like the schematic hierarchy in 4.2. Notice that the number of distinct cells in paradigms of this type equals $n+1$. Correspondingly, the number of total affix arrangements will be $B_{(n+1)}$.

Running my program using the feature systems with the maximum number of dependencies I obtained the resuts shown in table 4.4.

We see that in feature systems with many inter-dependencies among features, the proportion of overlapping homonymy is even higher than in systems with fully

Table 4.4: Overlapping homonymy in systems with many dependencies

feat. n	cells in a paradigm $n+1$	# of parad. with overlaps ?	p of parad. with overlaps $?/B_{n+1}$
2	3	0	0/5=0
3	4	2	2/15=.13
4	5	15	15/52=.29
5	6	91	91/204=.44
6	7	523	523/877=.60

independent features (if we contrast paradigms of the same size, not paradigms defined over the same number of features). For example, a 7 cell paradigm over 6 dependent features allows a much greater proportion of possible overlapping arrangements than an 8 cell paradigm with independent features (0.6 vs. 0.33). I tried a few other hypothetical systems with fewer dependencies, and in all of them the expected proportion of overlapping homonymy grows extremely fast as a function of paradigm size.

Thus, I conclude that if affix arrangement were completely arbitrary, most paradigms would involve at least some cases of homonymy; and if the universal feature inventory included more than 6 features (which it surely must), most of the paradigms with homonymy would include some cases of overlapps.

4.2 The underlying structure of agreement features

The choice of features and their organization is crucial in the analysis of inflectional identity because it determines to a large degree which affixes are syncretic or homonymous. For example, depending on how one construes the person features, consistent non-distinction between 1st and 2nd person can be viewed either as homonymy of the two (out of 3) independent feature values, or simple unin-

flectedness (irrelevance) of a single morphological feature in the language (e.g. the feature [± speaker]).

There is however no general agreement about what the underlying feature structure is for inflectional categories such as person and number. When it comes to person features, there are several proposals in the literature that differ with respect to what they pick out as a natural class.

For instance, Anderson (1992) analyzes the person features in terms of the values $\pm you$, and $\pm me$. Under this analysis, 2nd and 3rd person form a natural class as the $[-me]$ category, and 1st and 3rd person form a natural class as the $[-you]$ category, while 1st and 2nd person don't share any features in common.

On the other hand, Harley and Ritter (2002) propose a feature hierarchy for the pronouns in which only 1st and 2nd person form a natural class (see figure 4.3).[6] They cite sources going back to Forchheimer (1953) and Benveniste (1971) who in some remarks suggest that 3rd person is not even a true person. The old insight that 1st and 2nd person form a natural class is based on a number of empirical observations. For example, some languages only have 1st and 2nd person pronouns and use demonstratives for the 3rd person pronouns. Third person agreement is often zero marked, while 1st and 2nd person agreement is overt. Pro-drop in some languages (such as Hebrew and Finish) may be restricted to 1st and 2nd person. Also in some languages with split-ergativity, the split is conditioned by the person features: nominative-accusative case marking is used with 1st and 2nd persons while ergative case marking is used with 3rd persons (cf. languages such as Duirbal, Pashto). Harley and Ritter (henceforth H&R) remark that at least since Jakobson (1971), the key distinction between 1st and

[6]The nodes in capital font identify three major subgroups of features. The PARTICIPANT node specifies person features, the INDIVIDUATION node specifies number (group, minimal, augmented) and class features. And finally, the CLASS node encodes gender and animacy specifications. The labels in bold represent default interpretations of the organizing feature.

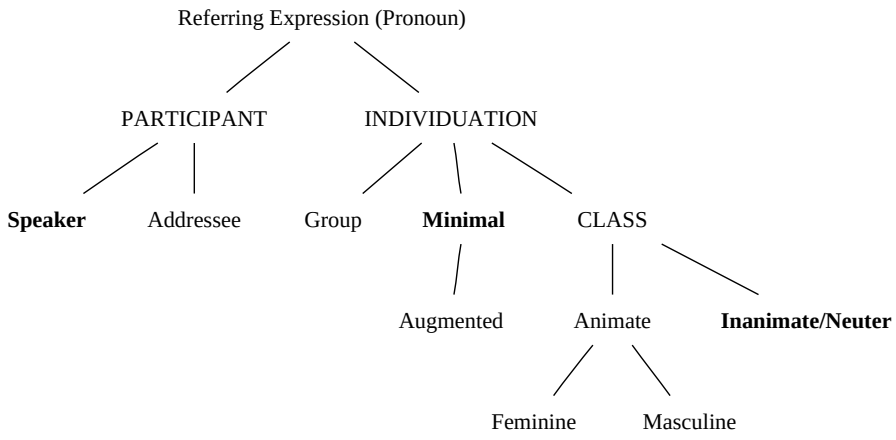

Figure 4.3: A morpho-syntactic feature geometry (Harley and Ritter, 2002)

2nd person, on one hand, and 3rd person, on the other, has been recognized as the difference between discourse dependent vs. discourse independent reference. That is, what "I" and "you" refer to depends on who is speaking or being addressed at the moment of speech, while the reference of the 3rd person is fixed.

H&R propose that the relevant person features are "participant" (in the speech event), "speaker", and "addressee". The last two features are dependent on the "participant" feature. Non-activation of the participant feature is viewed as a manifestation of the 3rd person. Activation of both "speaker" and "addressee" results in an inclusive 1st person marking. However, this theory of person features does not predict some of the attested person categories. More specifically, a number of languages (e.g., Hatam, Uradhi) distinguish situations that include a speaker and addressee vs. those that include the speaker, addressee and a 3rd person (Siewierska, 2004). This distinction is impossible to draw within a theory that treats 3rd person as a non-feature.

Cysouw (2001) proposes that combinations of 1+2+3 (speaker, addressee and other) persons be called "augmented inclusive," while the 1+2 is the regular

inclusive. These two combinations can stand in opposition to the exclusive 1st person (1+3). Cysouw suggests to use an 8 person system to describe person categories in both singular and non-singular numbers (instead of the the traditional 6 category system). This view is exemplified in the table below.

sing.	group	
	1+2	minimal inclusive
1	1+2+3	augmented inclusive
	1+3	exclusive
2	2+3 or 2+2	
3	3+3	

Since there is no agreed upon feature system for the person marking, I will remain somewhat agnostic about the precise details of such a system. I will, however, rely on insights and proposals of other linguists about what person categories form a natural class, which will be sufficient for the purposes of evaluating degrees and types of homonymy.

The natural classes with respect to person values (and their rough semantic descriptions) assumed here are summarized in table 4.5.

Table 4.5: Natural classes of person values

1+2	Participants in the speech event
1incl+1excl	including the speaker
1incl+2	including the addressee
2+3	excluding the speaker
1incl+1aug.incl	including the speaker and addressee

I have already discussed the reasons for grouping 1st and 2nd person into a natural class. There are also motivations for grouping 2nd and 3rd person into a natural class, which I will discuss shortly. On the other had, there is no strong empirical support for grouping 1st (excl.) and 3 person together, although both of these categories are logically connected as excluding the addressee. (Perhaps this

could be accounted for in a system where "addressee" was a dependent feature of the "+participant" node only.) The grouping of 1st person inclusive and exclusive forms is widely accepted as reflecting a natural class category realized syncretically in many languages (e.g., Indo-European languages). Also, since on practically all analyses of person, "inclusive" is analyzed as including the features of both the speaker and the addressee, I assume that 1st person inclusive also forms a natural class with the 2nd person. Minimal inclusive and augmented inclusive form a natural class according to Cysouw, although this grouping will be tangential for my purposes since no languages in my sample distinguished augmented inclusive.

Below, I go over some considerations in support of grouping 2nd and 3rd persons as a natural class. The view that 2nd and 3rd person form a natural class goes back to the old tradition in linguistics and anthropology (as discussed in Forchheimer (1953)). This grouping coincides with the intuitive (albeit egocentric) division of the world into us - 1st person - and everything else - 2nd and 3rd person.

This grouping is also supported by the facts about syncretism (including the data I discuss later). For instance, Baerman (2004) reports that many languages do not distinguish between 2nd and 3rd personal pronouns and 2nd and 3rd person agreement marking (either in the singular, or in the plural, or in both number categories). His data is based on the sample of languages considered in Cysouw (2001), with addition of a few languages from the syncretism database mentioned earlier.

Another indirect evidence that 2nd and 3rd person stand in opposition to the 1st person is the fact that in languages in which verbs agree for gender only in some persons, the division is either between the participants and non-participants

Table 4.6: Neutralization of person distinctions (based on Baerman, 2004)

	Num. of lang. in the sample	1p=3p	1p=2p	2p=3p
pronouns	18	1/18	5/18	12/18
verb agr.	27	5/27	9/27	13/27

with gender being marked on the 3rd, non-participant person (e.g., Ket, Harar Oromo, Swahili, Iraqw), or between 1st and non-1st persons with gender being marked on the 2nd and 3rd person (e.g., Hebrew, Olo, Beja). On the other hand, 1st and 3rd persons, to my knowledge, never pattern together in inflecting (or not inflecting) for gender to the exclusion of the 2nd person.

Turning to number, H&R assume that features such as "group" and "minimal" are involved in marking number distinctions (cf. the hierarchy in 4.3). Just like "speaker" and "addressee", these features can combine to form another category, in this case "dual." The intuition is that the most minimal group is a group of size 2. The feature "minimal" has a further dependent feature "augmented" that, when activated, can indicate "trial" or "paucal" numbers (paucal is usually used with small groups between 2 and 6 objects). In all examples given by H&R, the feature "augmented" always co-occurs with the feature "group". H&R do not discuss what happens when "minimal" and "augmented" are activated together without the feature "group" (which is logically possible given their hierarchy). I suggest a slight modification to their number hierarchy to make it more consistent (see figure 4.4).

According to my number hierarchy (depicted in figure 4.4), languages can be broken down into three types. Those in which only the feature "group" is relevant, distinguish between what we in English refer to as *singular* (-group) and *plural* (+group) numbers. The second type of languages also distinguish between minimal vs. non-minimal groups. Following H&R, I assume that depending on

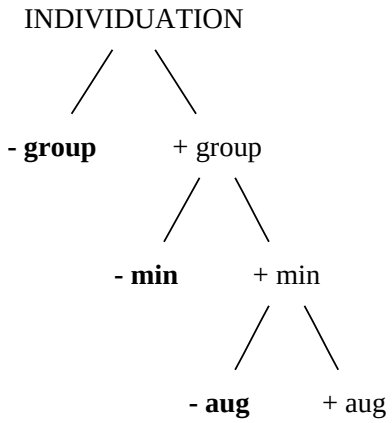

Figure 4.4: Number geometry

whether the language uses determinate or indeterminate way of counting in the sense of Corbett (2000),[7] this distinction could be instantiated either as "few" (paucal) vs. "many" (pl.) or as "two" (dual) vs. "more than 2" (pl.). That is, a minimal determinate group will be normally interpreted as a group of size two, while a minimal indeterminate group will be interpreted as a group of small size (usually 2-6). The third type of languages utilizes the feature "augmented" which refines the category "minimal" into minimal non-augmented and minimal augmented group. An augmented group has an additional member (or a few members) on top of the members included in the minimal group. This again could be either a division between "two" (dual) and "more than two, but still minimal" (paucal) or "two" (dual) and "three" (trial). Notice that this feature hierarchy accounts for Greenberg's universals that no language distinguishes trial without distinguishing dual and that no language distinguishes dual without distinguishing plural (Greenberg, 1963). This generalization follows from the general premise that if a language has a marked value of some feature, it also has

[7] "Determinate" refers to exact counting, such as one, two, three, etc. and "indeterminate" refers to approximate way of counting.

its unmarked value. As it is apparent from the number geometry I propose, dual is an unmarked counterpart of trial in languages that have the concept "augmented," i.e., dual is [-augmented] while trial in [+augmented]. Similarly, plural is an unmarked counterpart of dual in languages that have the concept "minimal group", i.e., plural is [-minimal] while dual is [+minimal]. The traditional terms *singular, dual, plural* are somewhat misleading as they can encompass different categories depending on which underlying features are activated in the given language. Thus, I will only use them when it does not lead to confusion, otherwise, I'll stick with the features *group, minimal* and *augmented*.

As for the analysis of gender and animacy, the features assumed in the H&R's analysis of the CLASS node (see figure 4.3) will suffice for my purposes. In this analysis, feminine and masculine form a natural class as the "animate" categories, in the opposition to the "inanimate"/neuter category. (However, for a more complete analysis, we would probably also like to distinguish grammatical gender within the inanimate category for languages in which inanimate nouns are marked for gender.)

4.3 Empirical data

Given the rough idea about expected frequencies (from section 4.1.2) and a concrete set of assumptions about the verbal agreement features, we now turn to evaluating the two hypotheses in (1) at the beginning of this chapter against the empirical data.

The data in my sample comes from 30 genetically and geographically diverse languages, all of which show some degree of form identity in the realization of agreement features. Most of these languages were taken from the University of

Surrey Syncretism Database. However, since this database was constructed with slightly different goals in mind, and since the examples it contained didn't always include all the relevant information, instead of querying the database I just used it for selecting languages with person/number syncretism so I could consult the relevant grammars for these languages.

All paradigms included in my sample encode subject agreement, or in a few languages, agreement with the most prominent argument.[8]

Agreement contrasts expressed in these languages include person (1st incl. or excl./2nd/3rd), number (sg/pl/du/trial), nominal class (including animacy, gender and some other class distinctions), degree of politeness, and switch reference. Occasionally, other verbal features are expressed cumulatively with agreement features, such as tense, modality, conjugation class, etc. A single language might have a different pattern of agreement across different tenses, mood, aspect, conjugations and other distinctions. Thus, a single language often contributed several different paradigm types to the sample. The total number of paradigms came to 93. Paradigms that were isomorphic to each other, i.e., that had the same inflectional pattern but different affixes, were not counted more than once. Homonymy and syncretism sometimes occur across these different paradigms; however, to narrow down the scope of inquiry, I will only focus on identical realization of feature values *within a single agreement paradigm.*

The information about the languages included in the sample and the number of paradigms contributed by each language are summarized in table 4.7.

[8]In some languages, verbs agree with whatever argument is higher on some hierarchy of prominence, usually the person hierarchy 1 >> 2 >> 3.

Table 4.7: Language sample

	Language	Family	Region	Num. of paradigms
1.	Aleut	Eskimo-Aleut	North America	1
2.	Amahuaca	Panoan	South America	3
3.	Amele	Trans New-Guinea	New-Guinea	4
4.	Atakapa	Gulf	North America	1
5.	Bagirmi	Nilo Saharan	North Africa	3
6.	Beja	Afro-Asiatic	North Africa	1
7.	Bulgarian	Indo-European	East Europe	2
8.	Burarra	Australian	Australia	1
9.	Burushaski	Isolate	South Asia	9
10.	Canelo-Craho	Macro-Ge	South America	2
11.	Carib	Carib	South America	2
12.	Cayuvava	Isolate	South America	2
13.	Chinantec	Oto-Manguean	Central America	7
14.	Daga	Trans New-Guinea	New-Guinea	12
15.	Dargwa (Icari)	North Caucasian	South Russia	4
16.	Diola-Fogni	Niger-Kongo	West-Africa	2
17.	French	Indo-European	Europe	5
18.	Harar Oromo	Afro-Asiatic	North Africa	3
29.	Hebrew	Afro-Asiatic	Middle East	3
20.	Hayu	Sino-Tibetan	South-Asia	2
21.	Hindi	Indo-European	South Asia	5
22.	Ibibio	Niger-Kongo	West-Africa	4
23.	Ket	Yeniseian	Asia	3
24.	Kiwai	Trans New-Guinea	New Guinea	1
25.	Krongo	Nilo-Saharan	North Africa	1
26.	Kwamera	Austronesian	South-East Asia	1
27.	Ngarinjin	Australian	Australia	1
28.	Olo	Torricelli	New-Guinea	1
29.	Rongpo	Sino-Tibetan	South Asia	6
30.	Teribe	Chibchan	Central America	1

4.3.1 Observed frequency of natural class syncretism

In this section, I evaluate the first empirical hypothesis in question, namely that inflectional paradigms avoid homonymy, but might include instances of natural class syncretism. Bear in mind that the counts presented here will underestimate the proportion of paradigms with no homonymy, as the selection criteria for the languages in the sample required that they contain some instances of syncretism or homonymy. In other words, this sample does not include languages in which subject agreement is marked unambiguously. If it turns out that even in such a biased sample most cases of ambiguity are due to natural class syncretism, this would be a strong confirmation for the hypothesis under consideration.

To count instances of natural class syncretism of person I relied on the natural classes identified in table 4.5. For number, I used the hierarchy in figure 4.4 in which each non-terminal node corresponds to a grouping of several categories into a natural class. For gender and animacy, I used H&R's hierarchy of CLASS features.

Form identity was identified as natural class syncretism if all paradigm cells that were occupied by the identical morph formed a natural class and no other morph occured within these cells.

Out of 30 languages in my sample, 25 had natural class syncretism in one or more of their paradigms. Out of the total 93 paradigms, 41 contained only natural class syncretism, and 21 contained natural class syncretism in addition to other kinds of inflectional identity. Additionally 7 paradigms contained no instances of form ambiguity at all. In other words, about half of the paradigms in the sample have no homonymy (41+7)/93 ($\approx 52\%$). The language by language breakdown showing the number of paradigms with natural class syncretism can be found the table 4.8.

Table 4.8: Number of paradigms with natural class syncretism and no homonymy

Language	# of parad. with only natural class syncretism	# of parad. with natural class syncretism and other types of syncretism
Amahuaca	2	0
Amele	3	0
Atakapa	1	0
Bagirmi	0	1
Beja	0	1
Bulgarian	1	0
Burushaki	2	5
Carib	1	0
Cayuvava	1	0
Chinantec	5	0
Daga	5	4
Dargwa	2	1
French	2	0
Harar Oromo	2	1
Hayu	1	1
Hebrew	2	1
Hindi	4	1
Ibibio	1	0
Ket	1	1
Kiwai	1	0
Krongo	0	1
Kwamera	0	1
Ngarinjin	1	0
Olo	0	1
Rongpo	3	1

In table 4.9 I take a more detailed look at the person syncretism. This table reports only those instances of form identity in which different person values are fully syncretic across one or more numbers. It does not include cases where person values are realized by the same affix across a partial range of contexts, such as when 1sg = 2sg = 2pl ≠ 1pl. The upper half of the table corresponds to instances of natural class syncretism, which predominate over all other types of identity among person values. Among the natural class patterns, homonymy between 2nd and 3rd person inflectional markers seems to be particularly common, consistent with the facts reported by Baerman (2004) (see table 4.6) and Baerman et al. (2005).

Table 4.9: Person syncretism in more detail

form ambiguity	sg	du	pl	number independent
1=2	Amele, Daga, Burushaski		Bagirmi, Rongpo	Amahuaca
2=3	Hindi, Atakapa, Hayu, Bulgarian, French,Rongpo, Amele, Ibibio	Hayu,Amele	Amele,Oromo, Carib,Chinantec, Burushaski, Ibibio	Chinantec, Amahuaca, Kiwai
1inc.=1excl	–		Cayuvava	
1=2=3	French,Dargwa		Burushaski, Rongpo	Hindi,Rongpo Chinantec
1incl=2=3		Hayu		
1incl=3		Kwamera		
1excl=3				Canelo-Craho
1excl=2			Burarra	

As for number syncretism, there were no languages where singular markers were always identical to the plural markers to the exclusion of dual or trial. Similarly, there were no languages where dual or trial markers were always identical to the singular markers to the exclusion of plural. In majority of the languages, no distinctions were made among the [+group] numbers, so the only opposition was

singular vs. plural. Seven languages made a distinction between "minimal group" vs. "non-minimal group" and out of these seven two distinguished dual (non augmented) vs. trial (augmented) minimal groups. In other words, the hierarchy in 4.4 makes correct predictions about attested patterns of number marking. All instances of inflectional identity in which one or several different number categories were fully identical across some person category presented cases of natural class syncretism (for more details see table 4.10)

Table 4.10: Number syncretism in more detail

form identity	in some persons	in all persons
−min.group = +min.group	Amele, Olo	
−group = −min.group = +min.group	Kwamera	
−group = +group in lang. with no min.group	Burushaski, Rongpo, Chinantec, Daga, Ket, Icari Dargwa	Chinantec, Icari Dargwa

Similarly, most cases of form identity of gender markers across other categories were instances of natural class syncretism. They all involved full neutralization of gender contrasts, usually in the plural and in the 1st and/or 2nd persons. Languages that had paradigms with gender syncretism are Beja, Hindi, Hebrew, Harar Oromo, Krongo, Ket, Olo and Ngarinjin.

There were also a few paradigms in the sample in which absolutely all distinctions were neutralized, i.e., the same form was used with all possible person/number/gender combinations.

Overall, we see that in about half of the verbal agreement paradigms, the instances of form identity are due to natural class syncretism. By itself, this does not seem to support the hypothesis that homonymy is dispreferred. But remember that our sample underestimates the proportion of non-homonymous paradigms because the Syncretism database that this sample is based on ex-

cludes languages that don't exhibit form-identity in the first place (or that don't inflect for subject agreement). To see how significant of an underestimation this is, consider that out of the 197 languages in the World Atlas of Language Structures 57 don't show verbal agreement for the subject (and thus, don't have any homonymy in agreement paradigms), 60 contain some syncretism or homonymy in verbal agreement, and 80 mark person/number unambiguously (Haspelmath et al., 2005) (see figure 4.5). So, based on this information we can already determine that at least $\approx 70\%$ of languages have no homonymy in their verbal agreement paradigms (including those which simply don't distinguish any agreement categories). As we've seen, the remaining languages, those that have some syncretism and/or homonymy, are such that 52% of their paradigms (according to my estimates) don't contain homonymy but only natural class syncretism. Thus, we can be reasonably sure that, at least with respect to verbal agreement paradigms, non-homonymous mappings predominate in inflectional systems, i.e., about $70\% + (.52 * 30\%) \approx 85\%$ of agreement paradigms contain no homonymy (these estimations assume that on average languages in my sample are not significantly different from other languages in the number of agreement paradigms they have).

Also recall that paradigms with no homonymy are expected to be rather rare by pure chance (given certain assumptions about the feature system discussed in section 4.1). This fact, together with the high observed frequency of non-ambiguous paradigms, support the first hypothesis in (1): namely, that languages avoid homonymy.

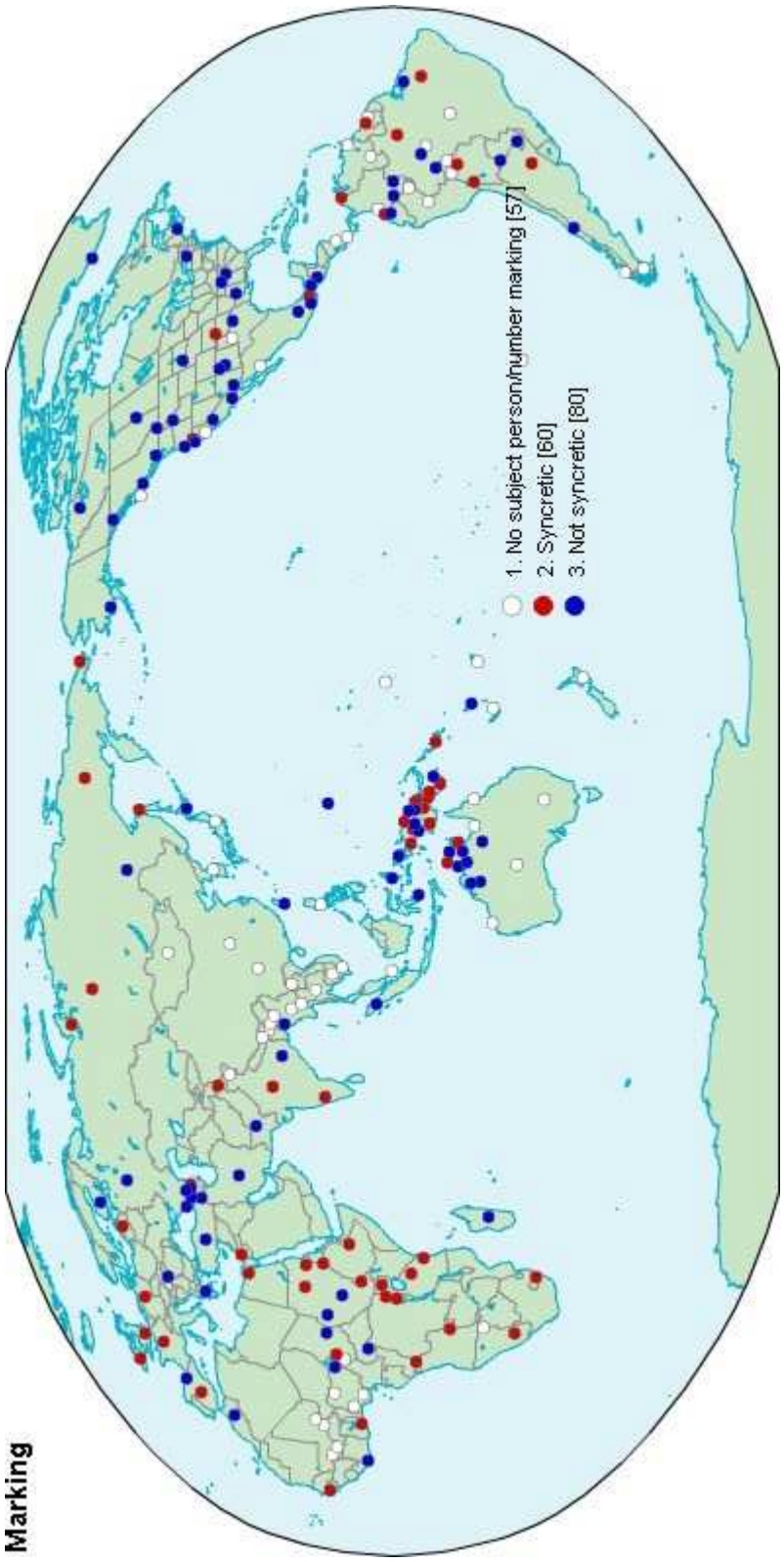

Figure 4.5: Person-number syncretism from the World Atlas of Language Structures (Haspelmath, 2005)

4.3.2 Observed frequency of elsewhere and overlapping homonymy

In the last section we established that about half of all paradigms in the sample contain no homonymy, since all examples of form identity in these paradigms are due to natural class syncretism. In this section we will evaluate the second hypothesis in (1), namely that when homonymy does occur, it is rarely overlapping (and therefore can usually be described with default reasoning).

To count the attested frequency of overlapping patterns, I again rely on the natural classes of inflectional categories identified in table 4.5, and the definition of overlapping homonymy. Since the invariant features of a morph could be viewed as the intersection of the smallest set of cells forming a natural class occupied by that morph, we can use the definition of overlapping homonymy as follows. Two morphs are overlapping if the smallest natural class containing one morph also contains the other and vice versa.

There were only 9 instances of overlapping homonymy in my sample, coming from 9 different languages. One such example comes from the Daga paradigm of past tense medial suffixes discussed in chapter 3 in section 3.1. Observe that in this paradigm, the suffix -*a* occurs within the domain that constitutes a natural class (1/2 person), but this domain also contains a suffix -*i*. At the same time, the smallest natural class containing all occurrences of *i* is the whole past tense paradigm, but it also contains the suffix *a*. For another example of an overlapping pattern, consider the following class prefixes in a Caucasian language spoken in Dagestan, Icari Dargwa. In this language the class prefixes on the verbs (indicating gender and animacy) differ depending on the number and sometimes person (Sumbatova and Mutalov, 2003).

The distribution of these prefixes is summarized in table 4.11.

Table 4.11: Class agreement prefixes in Icari Dargwa (Sumbatova & Mulatov, 2003)

		singular	plural
1/2p.	masc	-w-/-Ø-	-d-/-t-
1/2p.	fem	-r-	-d-/-t-
3p.	masc	-w-/-Ø-	-b-
	fem	-r-	-b-
	inanim	-b-	-d-/-t-

Observe that in the above paradigm the distribution of the morphs *-d-/-t-* (which are phonologically conditioned allomorphs) overlaps with the distribution of the morph *-b-* such that neither of them is more specific than the other, or can be said to block the other. *-d-/-t-* occur in the 3 person (the smallest natural class for *-b-*), and *-b-* occurs in the plural (the smallest natural class for *-d-/-t-*). This is the hallmark of overlapping homonymy.

Also, none of the 9 paradigms with overlapping homonymy involved more than 2 overlapping homophones. That is, even in paradigms with overlapping homonymy, majority of morphemes had either a non-ambiguous or an elsewhere distribution.

The total proportion of overlapping homonymy in the sample of 93 paradigms is 9/93, or about 10%. Since this sample is biased towards homonymous paradigms, the actual frequency of this type of homonymy will be significantly less than 10%. Additionally, the infrequent occurrence of overlaps cannot be attributed to the fact that such patterns are expected to be rare, since, as we have seen in section 4.1.3 they are likely to be very frequent by chance. However, it is possible that the number of overlapping patterns would go up if we consider larger paradigms spanning conjugation classes, tenses, etc. But at the same time, homonymy in such larger paradigms is more likely to be disambiguated by contextual and distributional factors (some of which I discussed at the beginning of this chapter). In

Table 4.12: Breakdown of paradigm types

Paradigm type	Num. (out of 93)
No form identity	7
Nat.class syncretism only	41
Elsewhere homonymy only	19
Overlapping homonymy only	5
Mixed homonymy/syncretism	21

other words, it seems to me that homonymy should be most problematic within the smaller domains, e.g. within a single subclass of affixes that occur in the same slot, and have similar distributional patterns belonging to the same inflectional class.[9] This reasoning was one of the rationals for limiting the window of investigation to agreement sub-paradigms only.

As for the "elsewhere" homonymy, it occured in 19 paradigms with no overlapping patterns or natural class syncretism. Elsewhere homonymy also often co-occurred with natural-class syncretism. The complete breakdown of the paradigm types discoved in the sample is shown in table 4.12.

Out of the 21 paradigms with mixed homonymy/syncretism types, 17 had natural class syncretism and elsewhere homonymy and 4 involved overlapping homonymy (together with natural class syncretism or elsewhere homonymy). Thus, overall the number of paradigms, that could be accounted for with underspecification and defaults was 84 of 93 paradigms, i.e., 90% of paradigms in the sample (and even more in the larger unbiased sample of languages). In short, the second hypothesis under investigation is also confirmed: among attested patterns of homonymy, overlapping patterns are significantly rarer than the elsewhere patterns within the window of a single agreement paradigm.

[9]The learner I propose here will not make a distinction between small vs. large domains, however I believe that a more realistic learner will begin generalizing within smaller domains or sub-paradigms extending the generalizations further only when they don't lead to problematic cases.

As a side note, free variation also appeared to be rare in the paradigms that I looked at. Whenever the grammars reported allomorphs, they were normally assigned to different inflectional classes, or there was some remark about conditioning factors. There were only three instances of allomorphy in verbal paradigms that represent clear examples of free variation. These examples come from Bagirmi, Daga, and Rongpo. For instance, in Rongpo many paradigms contain variants for 3 person singular, and sometimes other person/number combinations as well (see the paradigm for the copula "be" below).

Table 4.13: The Rongpo verb "be", present tense

1p.sg	hinki
2p.sg	hini or hin
3p.sg	hini or yã
1p.pl	hini
2p.pl	hini
3p.pl	hini or yã

To summarize this chapter, the available evidence leads us to believe that the hypotheses about statistical restrictions on homonymy in inflection formulated at the beginning of this chapter are true.

This raises the question of why languages have these tendencies, or why they prefer certain types of form-meaning mappings over others? The answer suggested in this dissertation is that non-homonymous mappings are particularly easy to learn, while overlapping mappings are particularly hard, with the elsewhere patterns lying somewhere in between. The learning model developed in the next chapter captures this intuition and makes further predictions about the shape of the resulting grammars and patterns of overgeneralization occurring during the learning phase.

CHAPTER 5

Learning

5.1 Introduction

In chapter 2, we saw that speakers possess a mental lexicon which includes representations of inflectional morphemes. We believe that such lexicons, or mental dictionaries where phonological units are associated with semantic and syntactic information, are necessary for speakers to be able to generate larger expressions such as words, phrases and sentences. In addition, we assume that the lexicon is minimal in the sense discussed in section 2.1.2.

The question we now face is: how can such a lexicon be learned? We know that speakers are not exposed to morphemes in isolation and they are not explicitly told what words mean. Instead, they hear fluent speech in different situations. Thus, the learner should be able to reason across such situations to extract features that are relevant to the speech signal and to find a mapping between such features and the units of form. In chapter 2, I discussed an intuitive cross-situational approach to this mapping problem that provides a first rough idea for lexical acquisition. This approach will form the core of the more sophisticated inflectional learner proposed here.

Since the problem of lexical acquisition is quite complex in its entirety, and since I am mainly focusing on one aspect of it - learning form-meaning correspondences in a way that fits the frequency patterns of form-meaning mappings

- I make a number of idealizations which I will point out as we go along. One of the big idealizations due to the focus on inflectional morphology is that the learner's input consists not of sentences, but rather of inflectional sub-sequences of words of the same part of speech. The inflectional strings segmented into morphs are paired with featural representations of the semantic context in the form of semantic feature values (refer to chapter 1 section 1.4 for a more detailed discussion of the input to the learner). Other kinds of idealizations pertaining to more specific morphological phenomena are discussed later in this chapter.

We can think of the learning space as organized into increasingly larger subsets according to a particular complexity hierarchy. In our case, this hierarchy is based on the empirically supported hypothesis that the overlapping patterns are more complex than the elsewhere patterns, which are in turn more complex than the one-to-one patterns. This is demonstrated in figure 5.1. The smallest subset, H1, includes paradigms with 1-1 form-meaning mappings, a slightly larger subset, H2, also includes paradigms that can be dealt with by default reasoning. Finally, the largest subset, (H3), includes all types of ambiguous and non-ambiguous mappings (except for paradigms with free variation, which are excluded from consideration). Assuming this kind of structured hypothesis space motivated by empirically grounded complexity considerations is similar to the idea used in Structural Risk Minimization framework for statistical learning (Vapnik, 2000).

The General Homonymy learner I propose at the end of this chapter will be able to learn any language from the set H3, although it will be biased to first select hypotheses from H1 and then from H2. That is, it will move to H3 only as a last resort. I build up to this learner by first considering simpler learners for the spaces H1 and H2.

In the remainder of this introductory section I do three things: I briefly

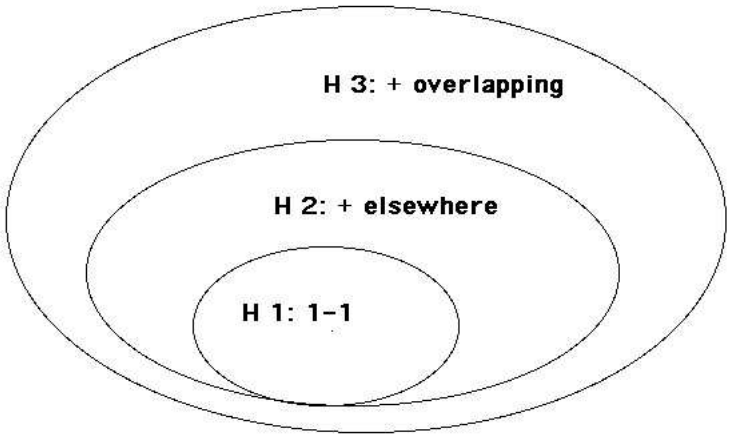

Figure 5.1: The hypothesis space based on the proposed complexity criteria

introduce some some basic concepts used in the work on formal learning, defend the view that a psychologically plausible learner must generalize, and discuss some previous computational work of Jeff Siskind that also relies on cross-situational learning to learn lexical meanings.

5.1.1 Setting the stage

Following the classical learning theory perspective (Gold, 1967; Blum and Blum, 1975), we will think of a *target language* as a set of *expressions*. In our case, the expressions are pairs of inflectional strings and *environments*, where a string is a sequence of morphs and an environment is a complete assignment of all universal inflectional features to some value. For example, if there are only 3 universal features {F1,F2,F3} and they are all binary, the following set of feature values constitutes an environment: $\{F1 : +; F2 : -; F3 : -\}$. I will sometimes refer to the target languages described above as *inflectional languages*.

So, a single expression is a string-environment pair. An infinite sequence of *all* expressions from a language is called a *text* (a text may contain more than one repetition of any expression). Learners are exposed to such positive texts and their job is to identify what language the text comes from. This is done by identifying a grammar that can generate the expressions of the language (in our case, the grammar is the lexicon). A learner is said to *converge* on a text if after a certain point it never changes its hypothesis. Such convergence counts as successful learning if the grammar that the learner converges on generates the language of the text. A learner successfully learns a language if it can successfully converge on every text for that language.

Before I discuss the grammars for the set of inflectional languages, I would like to make a few comments about some properties of the languages themselves. Sequences of inflectional morphemes are bounded in length - we don't see unbounded recursion in inflection (although perhaps one could argue that such recursion exists in derivation). Moreover, I assume that there is a finite bound on the semantic distinctions that can be marked inflectionally, and that there is a finite bound on the distinct morphs used to express these distinctions (i.e., there is no infinite synonymy). Given these three facts, we know that the target languages are finite.

This means that such languages could in principle be learned by a *memorizing* learner that simply records each new data-point it sees. Since the languages are finite, such a learner would eventually see all the data, and at that point it would have correctly converged on the target language.

In the next subsection, I discuss some reasons for why, such a simple memorizing learner will not do if we're trying to model human learning or even if we are simply trying to construct an efficient learning strategy. This discussion is

not meant as an argument against a view that anyone holds, but simply as a way of delimiting learning strategies and understanding the nature of the hypothesis space.

5.1.2 The need to generalize

First of all, a purely memorizing and non-generalizing morphological learner is not realistic as a model of human learning because wug-tests and overgeneralization errors made by children strongly suggest that people generalize even when learning in a finite domain.[1]

There are at least two other reasons for why a learner that simply adds each input to a memory stack is not plausible. First of all, such a learner would learn any finite pattern in exactly the same way, and make no predictions about the regularities in morphological lexicons. That is, a purely memorizing learner would just as easily learn languages where affix ordering was completely random in every new word and languages where the order was fixed, or languages with thousands of arbitrary inflectional classes and languages with no inflectional classes, languages with no homonymy, and languages with lots of homonymy. In other words, a memorizing learner would tell us absolutely nothing about why languages are the way they are, which goes against the premise that many properties of languages are to some extent determined by the constraints on language learning and language evolution.

Second, the sheer size of inflectional languages makes learning by memorization alone not feasible: there is little hope that a human learner will hear all

[1]This consideration by itself does not mean that a learner cannot memorize every input pair, as long as he/she is also generalizing. For instance, there are several memorizing-type models in which generalization is achieved via analogy or general ability to compute similarities over the memorized information (cf. Exemplar Models in phonology).

inflectional sequences in all possible contexts in which they could occur in her lifetime. A generalizing learner on the other hand, may converge on the target language after seeing only a subset of the data (depending on the number of irrelevant features).

To see how big inflectional languages can get, consider the following calculations. We can compute the upper and lower bounds on the size of an individual language as a function of the number of morphs m, the maximum allowed string length p (in terms of number of morphs), the number of universal semantic features n, and the number of feature values for every feature q. The number of maximum distinct strings of length p in a language is m^p, and the number of distinct environments is q^n. Assuming that any string can be associated with any environment and that every environment must be associated with some string, the maximum number of expressions in a language in which all words are of length p is $m^p * q^n$ (if every possible string occurs in every environment) and the absolute minimum number of expressions is $1 * q^n$ (if all environments are associated with a single string). The actual number of expressions in any language lies somewhere between these two extremes.

To get an idea of how big the lower bound on the number of expressions can get, suppose that a universal feature inventory consists of 50 binary independent features.[2] The number of environments in such a language is already an astro-

[2]It is likely that the actual inventory of features is at least this big. For instance, consider just some of the common inflectional distinctions that can be marked on the verbs: aspect, tense, mood, voice, transitivity, reflexivity, switch reference, direction of motion, subj. person, subj. number, subj. animacy, subj. gender, subj. social status, definiteness, subj. location, obj. person, obj. number, ..., indirect object number/person etc. Also consider the fact that many inflectional distinctions mentioned above have to be defined by several features, or defined by a feature with more than two values. For instance, we saw that the number contrasts could be defined with three different features: group, minimal and augmented. Similarly gender and person contrasts are often analyzed in terms of several features (e.g. "feminine"; "masculine"; "speaker"; "addressee", "participant in the speech event" etc). When it comes to tense, even if we assume that it constitutes a single feature, it will have a great number of values, such as

nomical figure: $2^{50} \approx 1.13 * 10^{15}$. Even if the actual number of environments is a million times smaller than this (for instance, due to dependencies among features that make certain combinations of feature values impossible), the total number of environments is still huge: $\approx 1.13 * 10^{9}$. For comparison, the number of seconds in 80 years is only $\approx 2.5 * 10^{8}$. So, even if a learner were exposed to a new input pair every second, 80 years would still not be enough to get through all of the possible environments.

Finally, the actual number of pairs between inflectional strings and environments (i.e., the expressions in the target language) also depends on the number of possible inflectional sequences. Some highly inflective languages, such as Turkish, have thousands and even millions of such sequences. For instance, consider the following statistical data based on Turkish corpora from Kurimo et al. (2006).

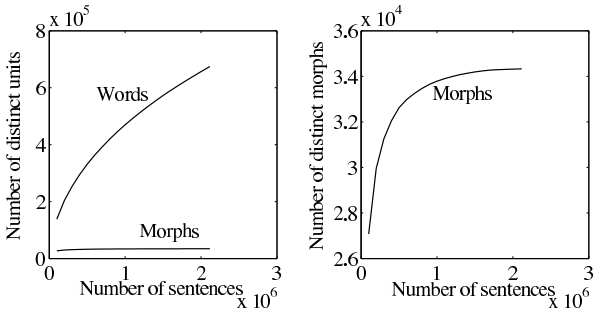

Figure 5.2: The growth of words and morphs in Turkish (Kurimo et al. 2006)

The graph on the left shows the growth of distinct word-forms and morphs in the corpus, while the graph on the right shows the same data for the morphs in more detail. We see that as the sample reaches 20 million sentences, new words continue to be encountered, while the number of morphs levels off. This graph is not specific to inflectional sequences, but it makes a general point that

present, past, future, distant past, distant future, immediate past, and so on. Same can be said about several other inflectional distinctions above.

in Turkish (and in most other languages) the number of morph sequences is generally considerably larger than the number of morphs.

Since, as we have seen, the number of environments and the number of morph sequences are already huge, the number of mappings between the two will be even bigger. The above considerations show that the purely memorizing, non-generalizing strategy is extremely inefficient and psychologically implausible.

5.1.3 The cross-situational learner of Siskind

The closest learning model that addresses a similar question as the one raised in this dissertation is the cross-situational learning algorithm proposed in Siskind (1996). This algorithm was designed for learning meanings of words from sentences paired with sets of *conceptual expressions*, such as CAUSE(**John**,GO(**ball**, TO(**John**))) for "John took the ball." The results of Siskind's work are somewhat hard to interpret since their presentation does not include a precise specification of the learning problem, proofs of convergence, or a rigorous discussion of the properties of the algorithm itself. From the general discussion and examples provided in the paper I can infer the following things.

Besides relying on cross-situational inference, Siskind's algorithm also takes advantage of the following assumptions: an empty set is not a possible meaning, and the meaning of an expression is equivalent to a multi-union of the meanings of its parts. The last assumption implies that (a) every symbol that is part of the meaning of a sentence must be attributed to some word in the sentence, and (b) words contribute non-overlapping parts to the meaning of a sentence. As I will discuss shortly, assumption (b) is not applicable in the domain of inflectional morphology, since sometimes several morphemes in the string can express some of the same features. Accordingly, we will assume that the meaning of a sequence

of inflectional morphs is equivalent to unions (rather than multi-unions) of the morphs' meanings. As for the assumption (a), while it also holds for inflectional sequences, it is not easy to take advantage of this fact since the learner does not know a priori what the exact meaning of the string is. Instead, it is presented with some superset of that meaning.

In Siskind's set-up there is also some uncertainty about the meaning of the sentence, but it is an uncertainty of a different kind. He assumes that the learners entertain several different hypotheses about the meanings of sentences, but that they have already converged on the relevant semantic symbols that can be used in the language. His assumption, when applied to the problem of learning inflectional sequences, is equivalent to a scenario in which upon hearing a word like *(jump)-s*, the learner entertains several disjoint hypotheses such as:

(1) (jump)-s:

a. JUMP, 3p. sg. pres.

b. JUMP, 3p. sg. past.

c. RUN, 1p. sg. future.

That is, the learner knows what conceptual symbols are appropriate in a given language, but it might incorrectly interpret the situation in several different ways. The first rule in Siskind's algorithm is designed to rule out the incorrect hypotheses before all other rules apply. It seems that in practice, the incorrect hypotheses can be identified quite quickly and that most of the time learning proceeds in a scenario in which an utterance is paired with a single meaning (at least, all of the examples considered in Siskind's paper assume a single meaning per utterance).[3] Most of the time the meaning that ends up being associated with an

[3]In cases in which an expression is globally ambiguous and a sentence could have several

expression after the first rule applies, corresponds to the correct meaning, though occasionally it may not (i.e., there is a possibility of noise).

I assume a somewhat different kind of uncertainty: although the learner can correctly interpret the semantic information contained in the environment (i.e., there is no noise), it does not yet know what semantic symbols are actually encoded by the speech signal. That is, I assume that upon hearing a sequence of morphs such as *(jump)-s*, the learner might be faced with the following set of semantic primitives:

(2) [JUMP; present, habitual, imperfective, intransitive, realis; subject: 3p.sg, anim, masc, respectful, located far away, etc.].

The learner has yet to determine that the morphology of the target language encodes only features of tense, mood, and subject's person and number, and how these features are lined up with the morphs.

In addressing the problem of homonymy and noise, Siskind's learner relies on several heuristics which are not perfect but in practice give good results (judging from the simulations on an artificial corpus of data). To detect homonymy or noise, Siskind's algorithm checks whether the set of *possible* meanings for a word (found through cross-situational intersections) is empty, or whether the set of *necessary* meanings for a word (found through some additional inferences) is not a subset of the *possible* meanings. He is able to take advantage of the latter heuristic because in circumstances in which a sentence is paired with the correct meaning, it is often possible to determine by process of elimination which meanings are *necessarily* part of some word. This inference, however, is not easy if irrelevant features are included in the mix or if a sentence is paired with an incorrect

meanings, one of the meanings is chosen based on a probabilistic metric.

meaning (which can lead to "corruptions" of the lexical entries). The other heuristic (observing whether the set of possible meanings becomes empty) works in many cases because open class homonyms are rarely used in situations that are very similar. For instance, the two meanings of the word *bank* ("financial institution" and a "river edge") are so distinct that if one takes an intersection of the two *conceptual expressions* associated with sentences including these two different uses of *bank* (which is how the set of possible meanings is calculated), one is likely to come up with an empty set. This heuristic, however, is not effective with inflectional homophones because they often share features in common. For instance, two different but homonymous morphemes may both be used in present singular environments, and their *possible* or invariant features will never be empty.

The learners I present here are not based on heuristics and can be proven to converge (with the Gold and PAC criteria for convergence[4]) on the languages they are designed to learn.

5.2 Assumptions about the hypothesis space

We can define the class of target languages in terms of constraints on the grammars that generate them. In this section I discuss such constraints and formally lay out the structure of the grammars for inflectional languages. Some of the constraints introduced here have not yet been discussed. Many of them are adopted to simplify or abstract away from various complicating morphological phenomena such as certain types of allomorphy, null morphs, variable affix ordering, etc.

Below is a high-level summary of the grammatical structure I assume. The specific properties of the grammars for inflectional languages will be first discussed

[4]PAC refers to the Probably Approximately Correct learning (cf. Valiant, 1984b; Anthony and Biggs, 1992)

informally and then formalized in section 5.3.

One of the main components of grammars for inflectional languages is the lexicon, which contains lexical items. Lexical items are morphemes consisting of a phonological realization (*a morph*, which could also be an empty string) and some semantico-grammatical representation. This representation essentially specifies how the morph is used. That is, it generally includes semantic properties that the morph expresses (sometimes referred to as properties of content) as well as properties of context, which limit the range of contexts in which the morpheme in question can be inserted. If a morpheme has no context specification, then it can co-occur with any other morpheme that is not competing to be realized in the same position. In here, I will largely ignore properties of context except for morphological properties (see 5.2.1).

In addition to the lexicon, the grammar also includes a slot template specifying what features might be expressed in what positions. Slots are often viewed as a reflection of the ordering of syntactic projections. The questions connected to morpheme ordering and consistency in feature realizations are taken up in section 5.2.2. Together, the lexicon, the slot template, and the principle of compositionality allow one to generate the language of string-environment pairs, where a string for our purposes is just a sequence of inflectional morphs. Recall that the term *environment* refers to a complete evaluation of a finite universal set of features corresponding to the properties that are true for the intended meaning. Each string is paired with some environment, namely an environment that comprises a superset of the meaning of the string. (Additionally, any context requirements on the morpheme combinations have to be satisfied.)[5]

[5]This picture of the grammar allows one to most naturally represent concatenative morphological systems, although I imagine that it could be extended to certain other systems as well with some modifications of the terms "slot" and "morph."

Besides the just mentioned requirements, the lexicon also obeys the minimality restrictions discussed in chapter 2, section 2.1.2. Recall that these minimality restrictions are part of the general preference for shorter representations among otherwise equivalent options.

In the next few subsections, I discuss some of the constraints on the inflectional languages I assume, as well as morphological phenomena that I will abstract away from for reasons of simplicity. These phenomena include allomorphy conditioned by phonological and lexical factors, variable morpheme ordering, portmanteau, and null morphemes.

5.2.1 Allomorphy and properties of context

First, I discuss the phenomenon of allomorphy, which will also give me a chance to motivate my assumption that the meaning of an inflectional string is better defined as a union of features of the constituent morphemes (rather than a multiunion). The learners I present later will not address the question of learning constraints on allomorphy, except for the morphologically conditioned allomorphy.

When different morphs express the same semantic features, but occur in complementary distribution, we call them *allomorphs*. Allomorphs are often likened to imperfect synonyms. The features of context mentioned in the previous section are introduced into the grammar to handle allomorphs, and ultimately to define constraints on legal morph combinations or *morphotactics*. Since allomorphs are constrained by their co-occurrence with other morphemes as discussed below, learning allomorphy results in learning morphotactics.

Morphologists usually distinguish three types of allomorphy: phonological, lexical and morphological (Haspelmath, 2002). Phonological allomorphs are con-

ditioned by co-occurring with morphemes that share some phonological properties in common. Examples of phonological allomorphs are the English inflectional endings *-[s], -[z]*, and *-[əz]* conditioned by the phonological properties of the preceding segment.

Lexical allomorphs are conditioned by co-occurrence with an arbitrary set of morphemes that have to be memorized. For instance, English speakers have to learn on the verb by verb basis which past participles take the suffix *-en* and which take the suffix *-ed*. Some declension/conjugation classes also present instances of lexically conditioned allomorphy (as long as stems that belong to the same class cannot be differentiated from other classes by some set of phonological or semantic properties).

Finally, morphological allomorphs are conditioned by co-occurrence with other morphemes expressing particular morpho-semantic feature(s). Recall the German verbal paradigm from chapter 2 (cf. table 2.1). In this paradigm, 3rd person singular is realized as *-e* in the past and as *-ø* in the present. Thus, we could say that the 3p.sg. allomorphs are conditioned by a tense feature, or by co-occurrence with morphemes that encode tense.

Morphological allomorphy could be easily learned if we ignore the distinction between the morphological features of content and context.[6] If we do this, we can then say that the German morpheme *-ø* has a meaning [3p.sg.,pres.] instead of [3p.sg] in the context [pres.]. I suppose the insistence on demarcating morphological features of context comes from avoiding situations where the same

[6]While, in principle, there is a distinction between the notions of content and context, this distinction is irrelevant in certain regards. For instance, from the point of view of a learner who is trying to learn the distribution of a particular morph, both content and context information present restrictions on this distribution: the content restricts the range of real-world situations in which a morph can be used, while the context restricts the range of combinatorial possibilities for the morph. In case of morphological allomorphy, the conditioning factors are semantic in nature and thus the distinction between content and context is further weakened.

feature could be expressed more than once. For example, the feature [+past] is already expressed in German by an independent tense morpheme *-t*. So, if we don't differentiate between features of content and context, we have to say that in words such as *spiel-t-ø* ("played", 3p.pres.) the property of tense is expressed twice - once by the tense morpheme, and once by the null agreement morpheme. But nothing prohibits us from saying this as long as the meaning of expressions is defined as a union rather that a multi-union, so that multiple expression of the same features does not change the meaning of the expression as a whole.

In fact, inflectional features are sometimes expressed redundantly either in different slots (a phenomenon called "extended exponence"), or in several words in a phrase that stand in the agreement relationship (e.g., pronominal clitics used in tandem with agreement morphology on the verbs). The meaning of the phrase in such cases is not different than if the agreement features were expressed only once. We also don't see the same inflectional morphemes apply to stems recursively, continuing to introduce more and more layers of meaning. These facts suggest that unions are the appropriate operation for calculating meanings of inflectional sequences. Therefore, it is possible to include what is often regarded as morphological features of context into the features of content without affecting the meaning of the expression as a whole (this does not apply to the contextual restrictions that are phonological or lexical in nature). The advantage of such an inclusion is that it provides an automatic way of disambiguating what is viewed as morphologically conditioned allomorphy.

In this thesis, I will not address the question of how phonological and lexical allomorphy is learned. The former task requires having a finer level of structure than what I have assumed (for instance, sequences of bundles of phonological features). At the end of this chapter, I will discuss some ideas about how the

latter task could be achieved by keeping track of co-occurrences of inflectional morphemes.

5.2.2 Slots and featural coherence

Another restriction on the target languages I adopt is the assumption that words can be analyzed as having a number of distinct positions or slots, and that each slot is designated for some set of inflectional features.[7] Morphs that occur in word positions corresponding to a particular slot realize some subset of the set of features appropriate for this slot. This assumption, which I call *featural coherence* (borrowing this term from Stump (2001), but using it in a slightly different way), is based on the idealized picture of empirical facts.

Many grammars incorporate the assumption of featural coherence by using slot templates in their exposition of the inflectional system. For example, Reh (1985) describes verbs in Krongo main clauses as having the following slot structure.

(3) Krongo Verb (Reh, 1985).

1	2	3	4	5	6	7	8
sbj.agr	tns/asp	stem	refl	deriv	pass	tempr	emph

I define featural coherence as follows: a morpheme belonging to some slot has to express a non-empty *subset* of the features appropriate for this slot. This allows some morphemes to be underspecified. For example, in Russian, verbal inflectional sequences that follow the stem can be broken down into two slots, assuming

[7] Although there are cross-linguistic tendencies with respect to affix ordering (Bybee, 1985; Trommer, 2003), there is still a considerable amount of variation and it is reasonable to assume that the order of affixes is part of what the learner has to learn about the morphological structure of her language.

that the first slot expresses features of tense and the second slot expresses features of person, number and gender. The second slot can also be said to express features of tense, or alternatively tense can be construed as a context feature (see table 5.1). Either way, there are two sets of morphologically conditioned allomorphs that occur in the second slot: one set is completely underspecified for person (these are agreement morphemes that co-occur with the past tense morpheme), and another set is completely underspecified for gender (agreement morphemes that co-occur with the present tense morpheme). The morphemes in slot 2 are featurally coherent in the sense that they all express some subset of the tense, person, number, and gender distinctions.

However, note that with no restrictions on the assignment of features to slots, the current definition of featural coherence is too weak in the sense that a whole set of universal features might be assigned to each slot, and then any morph can realize any feature in any position. However, the minimality restriction on the lexicons will rule out the possibility of assigning features to a slot if they are completely irrelevant for that slot (see formal definition of the minimality restriction in section 5.3). That is, I assume that if a feature is assigned to a slot, then it must have some effect on the strings of the language.

Table 5.1: Tense and agreement slots for some Russian verbs

slot O: stem	slot 1: tense	slot 2: tense,person,number,gender
govori- "speak"	-Ø- "present"	-u "1p.sg.present"
uchi- "teach"	-l- "past"	-ish "2p.sg.present"
etc.		-it "3p.sg.present"
		-im "1p.pl.present"
		-ite "2p.pl.present"
		-at "3p.pl.present"
		-Ø "masc.sg.past"
		-a "fem.sg.past"
		-o "neut.sg.past"
		-i "pl.past"

The advantages of a fixed morpheme order are great for learning. For instance, it provides for a possibility of bootstrapping. That is, once a learner observes a morph occurring in a certain position, and if she already knows what features are expressed in this position - she can zoom in on the meaning of the morph more quickly.

However, as I mentioned before, featural coherence and a fixed morpheme order are idealizations of the empirical facts. In some languages, the same properties are sometimes realized in different slots depending on the context. A point in case is Winnebago, where the instrumental prefix can either appear right before the verb root, or it can be sandwiched between the locative and the subject agreement prefixes. The reason for such alternations appears to be phonological in nature, since the position of the locative prefix depends on whether it contains a long or a short vowel (Hayes, 2005). Other types of factors that can affect affix ordering include a type of clause (e.g., main clause vs. subordinate clause), and morpho-semantic contexts (such as negative vs. non-negative sentence, past vs. present, etc.). Yet, changes in slot ordering usually affect only a few affixes and, as far as I know, there are no languages in which the order of inflections is completely arbitrary in every different word. Given that affix order is fixed at least within some morphological domain, I will limit my attention to such fixed orders only.

Another apparent exception to the concept of slots and featural coherence are the so-called *portmanteau* morphemes. Portmanteau morphemes can be analyzed as a fusion of several contiguous slots that are normally distinct in other strings in the language. This fusion results in a single portmanteau morpheme which expresses features of the contiguous slots simultaneously.[8]

[8]Sometimes non-concatenative realizations of inflection can create something very similar to a portmanteau morpheme. For instance, the irregular English form "sang" can be thought

Observe, however, that portmanteau morphemes cease to be exceptions to featural coherence if we extend the notion of coherence to apply at a higher level covering sequences of slots as well. In such case, we can always analyze portmanteau affixes as expressing features appropriate for positions spanned by the portmanteau morph.

Sometimes an analysis involving portmanteau morphemes can be restated using null morphs. To avoid these additional complications, I will focus on cases in which a single slot template can be posited, and every morph in every word can be said to express features of the slot corresponding to its linear position.

5.2.3 Null morphs

The last idealization discussed here is the assumption that null morphs have already been identified in the segmentation process and therefore are part of the input. This is a significant simplification because it is probably unrealistic to think that null morphs can be found based on phonological strings alone, with no semantic input (although it is sometimes possible to do so). In this section, I briefly outline some preliminary ideas about how null morphs can be discovered in tandem with learning form-meaning mappings. However, I leave the development of these ideas to future research.

Recall that null morphs are normally assumed as a representational device in order to describe non-overt realization of features in a minimal fashion. The assumption of null morphs is also convenient because it allows the analyst to maintain the idea that morphs are linearly ordered as specified by the slot-template.

If words are organized into slots, it should be possible to identify some null

of as a portmanteau morpheme expressing the features of two slots: the stem slot and the tense/agreement slot.

morphs during the segmentation process. For instance, the segmentation model of Goldsmith (2001) seems to do just that for languages that normally have a single inflectional morpheme following the stem. When a bare stem is discovered, one can automatically posit that it is followed by a null morph. In cases in which inflectional sequences are longer, the task of finding null morphs is harder, but it could possibly be achieved by using algorithms for finding the best alignment between maximally similar words that differ in length. For example, if we encounter two words of the form "a-b-c-d" and "a-b-d" and align them with each other so that they share as much common structure as possible, we would posit a null morph in the second word between "b" and "d". Undoubtably, this strategy is quite rough and it won't work in cases in which homonymy across different slots makes it more difficult to pinpoint the boundaries. For instance, when two adjacent slots can both contain the null morph and all overt morphs in these slots are homonymous, different strings (e.g., "a-Ø" and "Ø-a") would look identical on the surface (i.e., "a"). Morphotactic restrictions on the strings can further complicate the alignment procedure. Nevertheless, it is possible that finding minimal pairs and performing an alignment analysis could provide a first rough pass for identification of null morphs.

It is also possible to identify null morphs at the same time as finding minimal lexical mappings. I have pursued this idea previously, capitalizing on an observation that positing null morphs results in smaller lexicons. It is easy to see that after applying cross-situational intersections to strings that can be said to contain a null morph, the features that are expressed by the null morph will end up being associated with some other morphs in the string. For example, the English word "cat" will end up being associated with the representation [CAT; sg]. However, the lexicon will also contain a homonymous morpheme such as "cat" - [CAT] obtained from analyzing other words in which this stem occurs

in combination with overt affixes. Similarly, we will have pairs such as "dog" - [DOG] and "dog" - [DOG,sg]. After observing several homonymous lexical entries of this sort, the learner can minimize the lexicon by merging homonymous morphemes and positing a null morph expressing the feature *singular*.

However for simplicity of dealing with the problem of homonymy, I will assume that the input strings already have null morphs explicitly marked (this is also equivalent to an assumption that all morphemes are overt).

5.3 Definitions of the grammar and the language

We are now ready to define a language L of string-environment pairs (where the strings belong to the same part of speech). The expressions of L will serve as an input to the learner. The formal definitions below are helpful for better understanding the setup of the problem and for proving theorems about the learning algorithms.

Recall that the language is defined in terms of a grammar that can generate it. A grammar G for a language consists of an alphabet of morphs Σ, an alphabet of semantic properties F, a set of slots $[p] = \{1, ..., p\}$, a function Π that associates slots with sets of features, and a lexicon Lex consisting of lexical items, all discussed shortly.

Whenever I use the term *monomial* it refers to non-redundant consistent monomials (or combinations of feature values) as defined below.

(4) *Monomials*

For any $Y \subseteq F$, the non-redundant consistent monomials are $M(Y) = [Y \rightarrow V]$, where V is a set of feature values (in our case $\{+, -, n/a\}$).

That is, $M(Y)$ is the set of possibly partial assignments from Y to V.

The feature value "n/a" is used for those situations in which a feature cannot be evaluated because the presence of some other feature makes it non-applicable. For example, the number feature "minimal" does not apply in contexts in which the feature "group" is set to "−". Thus, the "n/a" feature is different from an underspecified feature (which is simply left out from the monomial). "n/a" is not compatible with "+" or "−".

The term *maximal monomial* refers to a monomial that constitutes a total function from Y to V.

(5) *Maximal Monomials*
$MM(Y)$ is the set of total assignments from Y to V.

(6) We will say that two monomials are consistent with each other if their union contains no contradictory features.

The affix slots are represented as a set of integers [p] ={1 ... p}. A function Π associates each slot $i \in [p]$ with a non-empty subset of the features $F_i \subseteq F$, more specifically, with elements of the power set of F, or $\mathcal{P}(F)$.

(7) $\Pi(i) = \{F_i | F_i \in (\mathcal{P}(F) - \emptyset)\}$

Note that the above definition excludes a situation in which some slot i is not associated with any features whatsoever, but slots are allowed to overlap in the features they express. Given the function Π we can define the Lexicon Lex in two steps. First, for each slot i in the template, I define a sub-lexicon of morphemes Lex_i that are said to belong to this slot. Then, the lexicon Lex can be viewed as

a union of all such sub-lexicons, i.e., $\bigcup_{i=1}^{p} Lex_i$.

Each sub-lexicon contains pairs whose first coordinate is a morph (an element of Σ), and whose second coordinate is a monomial of features. Additionally, sub-lexicons respect two requirements: featural coherence (see section 5.2.2) and minimality. The minimality requirement is in turn broken down into two parts: first we exclude useless (or irrelevant) features from being assigned to any feature slot; second, we require that morphemes be maximally underspecified (using strict underspecification). All of this is formulated below.

(8) $\forall \boldsymbol{i} \in [\boldsymbol{p}]$, $\boldsymbol{Lex_i} \subseteq (\boldsymbol{\Sigma} \times \boldsymbol{M(F)})$ such that:

a. featural coherence:

$\forall (m, v) \in Lex_i$, $v \in M(\Pi(i))$

b. minimality:

(i) no useless features:

$\forall f \in \Pi(i)$, $\exists (m, v) \in Lex_i$, such that $f \in v$.

(ii) strictly underspecify whenever possible:

$\neg\exists\ H = \{(m, v_1) \ldots (m, v_n)\} \in Lex_i$ for $n > 1$, such that $(v_1 \vee \ldots \vee v_n) \equiv \bigcap_{j=1}^{n} v_j$.

The first minimality restriction in (8),b,(i) rules out grammars in which there is a feature z assigned to some slot by the function Π, but no morph that occurs in this slot realizes z. The second minimality restriction (ii) rules out morphemes that are not maximally underspecified. This amounts to ruling out homonymous entries (lexical items that have the same first coordinate) if a single morpheme can be posited instead. This happens if the disjunction of the second coordinates of the homonymous lexical entries is logically equivalent ($\equiv$) to the intersection of their second coordinates. For example, $(\{p, -q, r\} \vee \{p, q, r\} \vee \{-p, q, r\} \vee$

$\{-p, -q, r\}) \equiv r = (\{p, -q, r\} \cap \{p, q, r\} \cap \{-p, q, r\} \cap \{-p, -q, r\})$.

The lexicon is defined as a union of sub-lexicons for each slot in the template [p]:

(9) $\quad Lex = \bigcup_{i=1}^{p} Lex_i$

To refer to the second coordinate v of a k-th slot morph m in the lexicon Lex (i.e., the meaning of m), I will use a notation $Mean(m, k)_{Lex}$. I will leave off the subscript whenever its clear what lexicon we are talking about.

With this definition of the grammar we can define the language L that consists of strings or morphs paired with complete assignments of all universal features, what I've been calling environments (or maximal monomials over F). In the definition below, s stands for the strings of morphs and e stands for the environments.

(10) $\quad L = \{(s, e) |\ \exists q \in \langle Lex_1 \times \ldots \times Lex_p \rangle,\ q = \langle (s_1, e_1) \ldots (s_p, e_p) \rangle$, such that

a. $\quad s = s_1 ... s_p$, (concatenation)

b. $\quad e \supseteq e_1 \cup \ldots \cup e_p$ (compositionality)

c. $\quad e \in MM(F)\ \}$

Observe that, because of the requirement that monomials be consistent (see (4)), the last condition above rules out combination of morphemes that contain contradictory features. Also note that the last requirement ensures that every string is paired with a maximal assignment of features. That is, every single feature from the universal feature set including the irrelevant features is part of the environment e.

The space of hypotheses entertained by the learner is restricted by the constraints on the grammars discussed above. Most of these constraints can be

interpreted as knowledge provided by UG. They include the assumption of compositionality, the assumption of featural coherence (morphemes that belong to the same slot express some subset of features assigned to this slot by Π), and the assumption of minimality. Recall, however, that the minimality assumption does not restrict the language, rather it is merely a restriction on representations: shorter representations are preferred over longer ones (see discussion in chapter 2).

The learner will be exposed to the language pairs (s,e) satisfying the conditions in (10) with respect to some lexicon *Lex* and a slot template Π. The job of the learner is to infer the identity of *Lex*. (Once *Lex* has been found, Π can be inferred from it.)

5.4 The No-Homonymy learner

The first learner I present is a simple cross-situational learner introduced informally in chapter 2. This learner will be shortly abandoned because it is ultimately inadequate for several reasons, one of which is its inability to handle same-slot homonymy. Nevertheless, it will serve as a good starting point for getting a concrete idea about how lexical mappings might be obtained from the string-environment pairs in a generalizing fashion.

The hypothesis space of this learner is restricted to languages with homonymy in the same slot. In such languages there is never a situation in which one morph is associated with several different meanings in the same sub-lexicon. This additional restriction on the grammar is formulated in (11).

(11) No homonymy within the same slot.

$\forall i \in [p]$, if $\exists(m, v) \in Lex_i$, then $\neg\exists(m, v') \in Lex_i$, where $v \neq v'$.

To learn the lexicon, the algorithm simply computes the invariant features for each different morph by taking intersections of the environments in which that morph occurs. (Identical morphs that belong to different slots are considered to be different.) This can be done incrementally and will result in a generalizing strategy. For example, upon seeing some morph in two different environments $[+f1, -f2, +f3]$ and $[-f1, -f2, -f3]$, the learner will infer that the meaning of this morph is $[-f2]$. This is a generalizing inference because it is performed without seeing the complete set of environments in which this morph can possibly occur (e.g., the environments $[+f1, -f2, -f3]$ and $[-f1, -f2, +f3]$). The cross-situational intersections (or underspecification) lead to discarding irrelevant features, and at the same time converging on the meanings of morphs.

5.4.1 The algorithm

The precise learning algorithm is presented in the pseudo-code format in Algorithm 1 below. Since we assume no portmanteau or null morphemes, the positions of each morph can be unambiguously identified by their linear order in the strings (we assign a subscript to each morph corresponding to its linear position in the string).

The algorithm takes a sequence of string-environment pairs of the form (s, e) as an input, and at each step returns an updated lexicon of morpheme pairs. For each morph s_i in the string s we check whether our current lexicon already contains a lexical entry in slot i, where s_i is the first member of the pair. Notice there could only be one such entry given our assumption of no homonymy within the same slot. If such lexical entry exists, then we update its current meaning by intersecting it with the environment e. If there is no such lexical entry, then we add a new entry (s_i, e) to the sub-lexicon Lex_i.

Algorithm 1 The No-Homonymy Learner

Input: a text T for the language L consisting of pairs (s, e)
Output: a set of sub-lexicons $Lex_1 \ldots Lex_p$ (all initially empty)

```
for all pairs (s, e) in Lg do
  for all s_i ∈ s do
    if ∃ (s_i, oldMeaning) in Lex_i then
      newMeaning ← (oldMeaning ⋂ e)
      replace (s_i, oldMeaning) with (s_i, newMeaning) in Lex_i
    else
      add (s_i, e) to Lex_i
    end if
  end for
end for
```

We can prove that the algorithm described above correctly converges on a lexicon that generates the target language in the current setting, (i.e., assuming languages have no homonymy within the same slot, no co-occurrence patterns besides the morphological ones, and that every word consists of a sequence of morphs of the same length where each morph expresses the feature of its word-slot.) Moreover, for languages that have more than one irrelevant feature this convergence will occur before all the text has been seen. These proofs are provided in the next sub-section (5.4.2).

The No-Homonymy learner helps us see more clearly that languages in which all homonymous mappings can be disambiguated by the linear position in the string are easy to learn. This learner relies on a simple, generalizing, incremental and memoryless strategy, that goes through the input one pair at a time, adjusts its current hypothesis, and discards the processed input pair. The intersective generalization strategy is safe in the absence of same-slot homonymy since it does not lead to incorrect predictions. This is a testimony to the fact that non-homonymous mappings are advantageous for the language learners. The fact that such mappings, including natural class syncretism, are rather a norm than

an exception as we have seen in the last chapter, lends support to a learner that incorporates cross-situational strategy as a generalization method.

Kobele et al. (2003) have a similar implementation of the cross-situational learner based on Siskind's work. Their learner is even simpler due to an assumption that the input consists of an unordered set of morphs paired with a set of sememes that correspond to the exact meaning of the utterances. The cross-situational learner I described in this section is slightly different in that (i) it takes some (albeit minimal) advantage of the morpheme order and is hence able to handle homonymy in different word slots; (ii) it learns from a superset of semantic meanings, which means that even seeing a monomorphemic word in isolation is not sufficient to converge on its meaning right away.

Another learner that relies on an intersectional strategy is the monomial learner (Valiant, 1984a). This learner starts with the hypothesis that every single literal of n features is in the concept $\langle u_1, \bar{u}_1, ...u_n, \bar{u}_n \rangle$, and then eliminates literals that are contradicted in the positive examples of the concept. The result is the same as if we take an intersection of the positive examples. Monomials defined over a finite number of attributes k are efficiently PAC learnable in the presence of irrelevant features (or attributes) from positive and negative examples (Dhagat and Hellerstein, 1994). The learner I presented here is like a monomial learner in that it learns in presence of irrelevant attributes; the difference is that my learner learns a set of concepts at once, rather than a single concept. The fact that all the concepts for my learner are formally distinct (there is no homonymy) helps keep the learning strategy simple.

In the next section, I elaborate on this simple learner and show how it can be extended to learn languages that include the elsewhere type of homonymy within the same slot.

5.4.2 Proofs

The casual reader can skip this section if he or she is bored by technical proofs and is eager to proceed to the next learner. The theorems proved here establish that the No-Homonymy learner converges on the right grammar for the languages it sets out to learn. Most of the notation used here has already been introduced in the previous section.

The main theorem of this section, theorem 2, proves that given a language L, the invariant features of any morph m in L (found by the intersective strategy) are equivalent to the meaning of m in some lexicon Lex that satisfies all the requirements in (8) and (11) and that generates L. To prove this, I rely on the definition of 'language' given in the previous section, as well as lemma 1 and theorem 1 below.

Recall that completely irrelevant features are those features that are not part of the lexical meaning of any morph.

Definition 1 (Irrelevant features) *A feature z is irrelevant iff there is no morpheme in Lex whose meaning includes any value of z.*

The next lemma (used in theorem 2) falls out of this definition of irrelevant features.

Lemma 1 *If a feature z is irrelevant, then for every pair $(s,e) \in L$, $\exists(s,e') \in L$ such that e differs from e' only in the value of the feature z.*

Proof. *Suppose that some feature $z \in F$ is irrelevant. That is, no morph in Lex is associated with a feature set containing any value of z. Since every environment is a total function of F, some value of z is included in every environment. Suppose we take an arbitrary pair $(s,e) \in L$ and derive a new environment e' by taking e and changing*

the value of z to any other value. This environment has to be associated with the string s in L, since z, being irrelevant, is not included in the meaning of any morpheme and so has no affect on the phonological realizations. Thus, the pair (s, e') *must be in L.*

The next theorem, which is also used in the main theorem 2, is necessitated by the fact that in presence of certain co-occurrence patterns, it is possible to have several equally minimal lexicons generating the same language. The main point of this theorem is to show that if a morph always occurs in environments that contain some semantic feature value, then this value is included as part of the morph's meaning in at least one of the several generatively equivalent lexicons.

Theorem 1 *If a language L is generated by a lexicon Lex such that*

1. *L satisfies all restrictions in (8) and the restriction on no homonymy*
2. *in the strings of L, some k-th slot morph m always co-occurs with one of the morphs* x_i *from some set* $X = \{x_1 \ldots x_n\}$ *where* $C = \bigcup_{x_i \in X} Mean(x_i, Lex)$*, and* $C \not\subseteq Mean_{Lex}(m, k)$

then a lexicon Lex' *equivalent to Lex in all respects except that* $C \subseteq Mean_{Lex'}(m, k)$ *generates the same language L.*

Proof. Assume that (1) and (2) above are true. And assume for contradiction that Lex and Lex' *don't generate the same language. Since the only difference between Lex and* Lex' *is in the meaning of the morpheme m, then the only difference between the languages they generate should concern expressions that contain m. Take all expressions in L generated by Lex that contain m in slot k. These are expressions of the form (s,e) where* $s_k = m$ *and where* $C \cup Mean(m, k)_{Lex} \subseteq e$ *(by compositionality and assumption 2).* Lex' *must generate exactly the same expressions, since* $Mean(m, k)_{Lex'} = C \cup Mean(m, k)_{Lex}$ *(by definition of* Lex' *above).* Lex' *cannot generate any other expressions that are not generated by Lex since it is identical to Lex*

in all other respects. By the same reasoning we can show that all expressions generated by Lex′ that involve m are also generated by Lex. Therefore Lex and Lex′ generate the same language.

Before we turn to the main theorem, one more definition is in order.

Definition 2 (Invariant features of a k-th slot morph m)

$$I(m,k) = \bigcap_{(w,e)\in L, m=w_k} e.$$

This definition basically says that invariant features for a particular morph (computed by the cross-situational learner) are equivalent to the intersection of all environments associated with this morph in the language. We are now ready for the main theorem which proves the convergence of the algorithm 1.

Theorem 2 *Suppose that a lexicon Lex (which satisfies all the relevant criteria in (8) and (11)) generates a language L. From this lexicon we can derive an equally minimal and generatively equivalent lexicon Lex′ following the recipe in the theorem 1. We show that for any k-th slot morpheme* $(m,v) \in Lex'_k$ *the following is true:*

$$I(m,k) = v$$

In other words, the invariant features of any morph m are equivalent to the meaning associated with the morph m by the sub-lexicon Lex'_k.

Proof. *We will show that for any k-th slot morpheme* $(m,v) \in Lex'$*:*

(a) $I(m,k) \supseteq v$*, and*

(b) $v \supseteq I(m,k)$

Part (a) is easy to show. Assume some morpheme $(m,v) \in Lex'_k$*. By def-n of the language in (10) and the assumption of no homonymy, any* $(s,e) \in L$ *where* $s_k = m$ *is such that* $v \subseteq e$*. Since I(m,k) =* $\bigcap_{(s,e)\in L, s_k=m} e$*, it follows that* $v \subseteq I(m)$*.*

Part (b) is a bit trickier. Take an arbitrary morph m that occurs in the k-th position in strings of L. Suppose that a feature value l is in $I(m,k)$. That is, l is in the semantic environment of every string containing m in the k-th slot. We will show that l is necessarily in v, for a morpheme $(m,v) \in Lex'_k$.

Claim 1: l cannot be a value of an irrelevant feature. If l was expressing an irrelevant feature, then by lemma 1 the set of expressions whose strings contained m would include pairs like (s,f) and (s,f') where $l \in f$, $l \notin f'$. But then, l could not be in $I(m,k)$, contrary to our assumption.

Claim 2: Since $l \in I(m,k)$, l must be expressing a feature that is part of the environment of every string containing m in the k-th slot. The feature values in an environment are either irrelevant or contributed by at least one of the morphemes in the string associated with that environment. Since l is not irrelevant, either (i) l is part of the meaning associated with m (ii) or it is part of the meaning of every morph in some set X, and m always co-occurs with some $x_i \in X$ in every string of L. In case (i), l is necessarily in v of any lexicon generating L (compositionality). In case (ii), l is included in v by definition of Lex' (see the theorem 1). Thus, l is necessarily in v for (m,v) in Lex'.

In short, we have proved that in the simplest scenario assumed for the first learner invariant features for all morphs in L are equivalent to the second co-ordinates of these morphs in one of the minimal lexicons generating the language L.

5.5 The Elsewhere learner

In this section, we provide a learner for the second hypothesis space H2, which excludes overlapping distributions but allows elsewhere homonymy. Accordingly, we remove the restriction adopted for the first learner, namely the restriction

that languages contain no homonymy within the same slot. However, we replace it with a different restriction, the one demanding that there be no overlapping morpheme distributions. This is in one sense a weaker restriction because it allows homonymy within the same slot as long as it is not overlapping. On the other hand, it is a stronger restriction because it rules out free variation. However, both of these consequences, absence of free variation and of overlapping homonymy, get us closer to the empirical facts since these types of patterns are rare in inflectional paradigms (as we have seen in the previous chapter). Thus, ruling out overlapping patterns, is a less severe simplification of the facts than ruling out same-slot homonymy.

Recall that we have previously defined an overlapping distribution relative to the notion of a paradigm, where a paradigm was viewed as all possible combinations of a set of features (cf. chapter 3 section 3.3). I repeat the definition of overlapping distributions here.

(12) A paradigm contains an **overlapping distribution** of morphemes, if at least two morphs in the paradigm "overlap", i.e.:

   a. their invariant features are consistent with each other, and
   b. each morph occurs in a cell that is consistent with the other morph's invariant features.

This definition can be used to determine the overlapping distribution in the language as a whole rather than in a particular paradigm. To do this we just extend the notion of a paradigm to encompass the set of all combinations of universal features. In other words, each individual environment associated with a word serves as a paradigm cell in the above definition.

Recall that the overlapping distribution encompasses precisely those patterns that cannot be described by the Blocking Principle (for examples of overlapping patterns consult chapter 3).

There are several ways to formalize a ban on overlapping distributions. We can either try to exclude overlaps from the lexicon by introducing homonymous lexical entries when necessary, or, in accord with the blocking proposals, we can assume a separate grammatical component that resolves competition among lexical items while maintaining a lexicon in which there is a single entry for each different morph. I follow the latter strategy, since it allows to formulate a target grammar that can be learned by a simple generalizing method that continues to make use of cross-situational intersections.

5.5.1 Formalizing blocking

In this subsection I modify the definition of the grammar presented in section 5.3 so as to rule out overlapping distributions. With this new restriction in place, we can proceed to provide a learning algorithm for the languages generated by the grammar defined here.

In addition to Π and Lex, we will have another grammatical component that I call BR (for blocking rules). Instead of assuming a single Blocking Principle based on some high level generalization, BR will explicitly record what morphs block each other. Given this set of blocking rules, one can later infer a higher level general Blocking Principle if such a principle indeed holds true for a given language. Just like Lex, BR will be a union of slot-specific components $BR_1 \ldots BR_p$. Each set of blocking rules BR_i will contain a set of morph pairs, in which the first morph in a pair *blocks* the second morph in the pair. The conditions in (13) specify under what circumstances the blocking rules are posited.

(13) $\forall i \in [p]$, if Lex_i contains two different morphemes (m, v) and (m', v') where v is consistent with v', then one of the following is true:

a. $(m, m') \in BR_i$

b. $(m', m) \in BR_i$

c. (m'', m) and $(m'', m') \in Br_i$, for a morpheme $(m'', v'') \in Lex_i$, such that $v'' \supseteq v \cup v'$.

The last condition (c) says that if two morphemes have consistent lexical specifications, it could be that neither of them blocks the other, but that some other third morpheme blocks both of them (for a schematic example of this pattern see figure 3.1 (b) in chapter 3).

The blocking relation is transitive: that is, if m blocks m' and m' blocks m'', then m also blocks m''. However, loops are disallowed:

(14) For any i, if $(m, m') \in BR_i$, then $(m', m) \notin BR_i$.

Also observe that the formulation of the BR component above allows for some grammars in which certain lexical items are never used. This can happen if a pair (m, m') is in BR_i, and $Mean(m, i) \subset Mean(m', i)$ (i.e., m is more general than m'). In such a case morph m' will always be blocked by m and so it would be a "useless" lexical item. However, as you will see, the learning algorithm will never propose such strange grammars and so we can ignore them. This fact is discussed in more detail later in regard to the Subset Theorem (page 133) which highlights the empirical vacuousness of the Subset Principle.

We will say that some i-slot morpheme is a *winner* with respect to some set of feature values Y if its meaning is a subset of Y and if it blocks all other i-slot

morphemes whose meanings are also subsets of Y.[9]

(15) $(m, v) \in Lex_i$ is an i-winner(Y) iff

a. $v \subseteq Y$

b. $\forall (m', v') \in Lex_i$ s.t. $v' \subseteq Y$, $(m, m') \in BR_i$.

Now, we can modify the definition of the language given previously in (10) to incorporate blocking as follows:

(16) $L = \{(s, e) |\ \exists q \in \langle Lex_1 \times \ldots Lex_p \rangle$, $q = \langle (s_1, e_1) \ldots (s_p, e_p) \rangle$, such that

a. $\forall i$, (s_i, e_i) = i-winner(e), (blocking and compositionality)

b. $s = s_1 ... s_p$, (concatenation)

c. $e \in MM(F)$ $\}$

Note that the compositionality restriction is now automatically encoded in the restriction that words consist of *winners* with respect to the features of the environment: it follows from the definition of winners that the features of the environment are a superset of the union of the features of the constituent morphemes.

With the addition of blocking, we can now describe the "elsewhere" distributions of morphemes without positing homonymous lexical entries. At the same time, the blocking component excludes overlapping patterns since they are impossible to generate using the restrictions on the grammar I described in this section.

[9] We can imagine some sets of feature values that are ineffable or for which no unique winners exist. This could happen either because there are no morphemes compatible with this set of features or because the grammar somehow fails to resolve the competition (although in my definition of blocking, I require that the competition always be resolved, but we can imagine relaxing this assumption).

5.5.2 The algorithm

The first learner we considered was based on a simple intersective strategy that easily learned languages with natural class syncretism and homonymy restricted to different word slots. However, this simple learner did not have a way of detecting homonymy within the same slot. The learner I present next will be able to do just that for a subset of the homonymous affix distributions, namely the elsewhere distributions. Hence, I call it the Elsewhere learner. This learner relies on the same generalization method as the previous learner, and thus it can still easily learn any non-ambiguous mapping including natural class syncretism.

Besides the cross-situational intersections, the Elsewhere learner includes an additional routine for detecting homonymy and correcting its hypothesis by adding blocking rules. The lexicon still contains only a single lexical entry per each morph, and each morph is still paired with its invariant features. However, some morphemes have an elsewhere distribution due to the fact that they can be blocked in particular environments by other morphemes. Algorithm 2 on page 130 shows an implementation of this learner.

The general strategy for this learner can be described as follows. As before, the algorithm runs through a text T for a language L one expression at a time, and incrementally calculates the invariant features for each morph. However, now after every cross-situational intersection, we check if this intersection leads to unresolved lexical competition (i.e., a situation in which several same-slot lexical entries are compatible with the same environments, and the current blocking rules don't resolve the competition). In case some morph m has *competitors*, for every one of them we determine whether it can be posited as a blocker or a blockee of m, or neither (if some other morph blocks moth m and its competitor). This can be determined trivially if the hypothesized invariant features

stand in a subset relation to each other because the more specific morph has to block the more general one (see Theorem 4). If they don't stand in the subset relation, then we can determine the direction of blocking only if we have already seen what morph occurs in an environment that is consistent with the currently hypothesized features of both competitors. If such evidence has not been seen yet, no blocking rules are added and overgeneralizations are predicted to persist until the disambiguating data is uncovered. For instance, this could happen if some morph a were associated with the features [f1,f2], a morph b were associated with features [f3,f4], and a morph c were associated with features [f1,f2,f3,f4]. The algorithm will correctly diagnose this case by adding appropriate blocking rules if c has already been seen. If there is no evidence one way or the other about what morph occurs in the environments that are consistent with both competitors, no blocking rules are added and the learner continues to overgeneralize until the relevant data is discovered and c is posited as a blocker of both a and b. Recall that by our assumption of no overlaps, if a blocks b the reverse is impossible. Thus, once the direction of blocking has been determined, it will not be contradicted by any future data.

This learner will succeed on the languages with no overlaps because (i) in absence of homonymy it works just like the no-homonymy learner, (ii) it has a capacity to correctly diagnose the presence of "elsewhere" homonymy (see Theorem 3), (iii) when elsewhere homonymy is detected it can determine the right blocking relationships (see Theorem 4 and the subsequent discussion).

If one wants to have a memoryless strategy (in which the learner does not have access to all input pairs seen so far), one could adopt the assumption that the text is *fat*. That is, every expression of the language occurs in it infinitely many times although with different probabilities (Osherson et al., 1986). With

Algorithm 2 The Elsewhere Learner

Input: a text T for the language L consisting of pairs (s, e)

Output: a lexicon Lex = $Lex_1 ... Lex_p$, a set of blocking rules $BR = BR_1 ... BR_p$.

Mem $\leftarrow \emptyset$
for all pairs (s, e) in T **do**
  Mem $\leftarrow$ add (s, e) to Mem
  for all $s_i \in s$ **do**
    if $(s_i, x) \in Lex_i$ **then**
      $newMeaning \leftarrow (\text{x} \bigcap e)$
    else
      $newMeaning \leftarrow e$
    end if
    $competitors \leftarrow \{(m, v) \in Lex_i \mid (m, v) \neq (m, oldMeaning)$, v is consistent with $newMeaning$ and $\neg\exists(s_i, m) \in BR_i\}$
    for all $(m, v) \in competitors$ **do**
      if $v \subset newMeaning$ **then**
        ensure (s_i, m) is in BR_i (add it if it's not there)
      else if $v \supset newMeaning$ **then**
        ensure (m, s_i) is in BR_i
      else if $\exists(s', e') \in Mem$, s.t. $(v \cup newMeaning) \subseteq e')$ and $s'_i = m$ **then**
        ensure (m, s_i) is in BR_i
      else if $\exists(s', e') \in Mem$, s.t. $(v \cup newMeaning) \subseteq e')$ and $s'_i = s_i$ **then**
        ensure (s_i, m) is in BR_i
      else if $\exists(s', e') \in Mem$, s.t. $(v \cup newMeaning) \subseteq e')$ and $s'_i = x \neq s_i \neq m$ **then**
        ensure (x, m) and (x, s'_i) are in BR_i
      else
        {Do nothing. There is no evidence.}
      end if
    end for
    addreplace $(s_i, newMeaning)$ in Lex_i {add it if it's not there, replace the old entry if there is one}
  end for
end for

this assumption, we are guaranteed that, no matter where in the text we currently are, we will always see the input pair that allows us to determine the direction of blocking. The algorithm 2 below is a simplified version of the previous algorithm that takes a fat text as an input.

Algorithm 3 The Memoryless Elsewhere Learner

Input: a **fat** text T for the language L consisting of pairs (s, e)
Output: a lexicon Lex $= Lex_1...Lex_p$, a set of blocking rules $BR = BR_1...BR_p$.

1: **for all** pairs (s, e) in T **do**
2: **for all** $s_i \in s$ **do**
3: **if** $(s_i, x) \in Lex_i$ **then**
4: $newMeaning \leftarrow (\text{x} \bigcap e)$
5: **else**
6: $newMeaning \leftarrow e$
7: **end if**
8: $competitors \leftarrow \{(m, v) \in Lex_i \mid (m, v) \neq (m, oldMeaning)$, v is consistent with $newMeaning$ and $\neg\exists(s_i, m) \in BR_i\}$
9: **for all** $(m, v) \in competitors$ **do**
10: **if** $v \subset newMeaning$ **then**
11: ensure (s_i, m) is in BR_i (add it if it's not there)
12: **else if** $v \supset newMeaning$ **then**
13: ensure (m, s_i) is in BR_i
14: **else if** $e \supseteq v \cup newMeaning$ **then**
15: ensure (s_i, m) is in BR_i
16: **else**
17: {Do nothing.}
18: **end if**
19: **end for**
20: addreplace $(s_i, newMeaning)$ in Lex_i {add it if it's not there, replace the old entry if there is one}
21: **end for**
22: **end for**

Notice that both of the Elsewhere learners above learn the lexical mappings and rule out irrelevant features in a generalizing fashion. The generalizations occur not only during intersections, but also in the process of positing blocking rules. More concretely, after observing a single instance in which some morph

a blocks another morph b, we conclude right away that a is always a blocker of b (even without yet knowing the exact meanings of a and b). This kind of generalization is safe given the restriction of no overlaps adopted in this section.

This Elsewhere learner captures the relative simplicity of elsewhere mappings compared to the overlapping mappings: in languages with no overlaps one can still use a relatively simple generalizing strategy based on cross-situational learning without backtracking or changing the adopted lexical items.

5.5.3 Theorems related to the Elsewhere learner

In what follows, I present a few theorems that prove that the above learners correctly detect presence of homonymy, and correctly determine the direction of blocking in case competitors are in a subset relationship. If competitors are not in the subset relationship, the blocking rules are always correctly determined since the algorithm simply waits to see which morph will occur in the environment compatible with both of the competitor's invariant features.

Let the term "currently invariant features" refer to the features derived at some intermediate stage of computing invariant features. The first theorem below shows that whenever currently invariant features of two same slot morphs become consistent, we can conclude that at least one of the morphs is a homophone.

Theorem 3 *Adopting the restriction on no overlaps (i.e., no free variation and no overlapping homonymy), we prove that whenever the currently invariant features of two same slot morphs become consistent at an intermediate learning stage, we can infer that one of the morphs is a homophone.*

Proof.

Suppose that the language L contains no free variation and no overlapping homonymy. Furthermore, suppose that in the process of applying the Elsewhere algorithm to the

language L, the currently invariant features of two same slot morphs a and b become consistent. Since the currently invariant features are supersets of the invariant features (this follows from the way invariant features are calculated), the invariant features of the two morphs must also be consistent (i.e. contain no contradictory feature values). There are two possibilities: (1) neither a or b are homophones (2) at least one of them is a homophone. In the first case, the invariant features of both morphs correspond to the necessary and sufficient features determining their distribution. These features are also equivalent to their lexical meaning in a lexicon for L (see theorem 2). In the absence of homonymy, two morphs with consistent lexical meanings are predicted to stand in free variation in some environments. However, this contradicts our assumption of no free variation. Therefore (2) must be true.

Thus, consistency can be used as an indicator of homonymy.

Theorem 4 shows how the ban on overlapping distributions also helps in inferring the blocking direction from the invariant features. Namely, if the currently invariant features of the two competitors stand in a subset relation to each other, we can immediately determine which one of them blocks the other.

Theorem 4 (The Subset Theorem) *If the intermediate lexicon contains an entry (m, v), and a new entry (m', v') is about to be added such that $v \subset v'$, we can immediately infer that $(m', m) \in BR$.*

Proof: Suppose an intermediate lexicon Lex obtained by the Elsewhere learner contains a morpheme (m, v), and another morpheme (m', v') (where $v \subset v'$)) is about to be added to Lex. Since $v \subset v'$, then v and v' are consistent and by (13) the blocking rules must include one of the following: (i) (m, m'), (ii) (m', m), or (iii) (m'', m) and (m'', m') for some morpheme (m'', v'') where $v'' \supseteq v \cup v'$.

Suppose that (i) above were true. That is, m blocked m'. Then, it would be impossible for m' to be expressed in any language pair, since m would win over m' with

respect to any environment in which m' could possibly occur. Therefore, we couldn't have derived the lexical entry (m, v) in the first place. The same could be said about option (iii). Since $v \subset v'$, then $v \cup v' = v'$, therefore $v'' \supseteq v'$, and neither m nor m' could ever win with respect to m''. By process of elimination (ii) must be true.

This theorem shows the empirical vacuousness of the Subset Principle (or the Elsewhere Condition based on specificity). The empirical data could never provide us with any evidence that the more general lexical items block the more specific ones. The opposite state of affairs is the only empirically observable option. Thus, the Subset Principle is inherently and logically build into reasoning with blocking; it is not a separate principle that makes empirical predictions. The same point is made by Prince and Smolensky with regard to OT constraints that are in a subset relationship (cf. "Pananini's Theorem on Constraint Ranking" Prince and Smolensky, 1993).

5.6 The General Homonymy learner

So far we have seen that a language in which strings of morphs are associated with maximal monomials over a finite number of features is easily learnable given particular affix distributions, namely distributions that exclude overlapping patterns. Nevertheless, such patterns are empirically attested, and therefore, we need our learner to be flexible enough to learn them. In this section, I get one step closer to this goal by showing how any pattern of homonymy including overlapping homonymy can be learned (keeping a restriction on no free variation constant).

The learner presented here is called a General Homonymy learner. It extends the elsewhere learning strategy to handle any type of form-meaning mapping

except for free variation. This learner continues to generalize non-monotonically and it matches the empirical facts regarding frequencies of homonymy patterns because it has the easiest time learning one-to-one mappings and the hardest time learning overlapping mappings.

The next subsections elaborate on the target grammars for this learner, the learning method and the predictions that this learner makes.

5.6.1 The learning space

Languages with unrestricted homonymy can be described by positing several lexical entries with the same first coordinate (something we have not done yet). Allowing an unbounded number of homonymous lexical entries in the lexicon makes the use of blocking rules, technically speaking, unnecessary. However, we will only posit homonymous lexical entries as a last resort, so reasoning with blocking will still be useful as a simple and efficient method for learning majority of homonymous mappings (i.e., the elsewhere mappings). Therefore, I will continue to rely on blocking since I aim to construct a learner that easily learns simple and well-attested patterns in contrast to more complex and infrequent patterns.

In the present scenario, the restrictions on the languages in (10) continue to hold, as well as the restrictions on the blocking rules discussed in section 5.5. However, now there are no additional restrictions on homonymy, although we still assume that free variation is ruled out. This assumption is spelled out below.

(17) $\forall Lex_i$, $1 \leq i \leq p$, there are no two *different* morphemes in Lex_i whose meanings are consistent and the blocking rules do not resolve the competition among them.

Since we are allowed to have several lexical entries with the same first coordinate, we need a way to distinguish them from each other. For this purpose, I use integers starting from 1 to whatever the highest number of homophonous morphemes may be. So, now the first coordinate of the lexical entries is a tuple with the first member being a morph and the second member - an integer. (As before, the second coordinate of lexical entries is a set of feature values, or a monomial).

(18) $Lex_i \subseteq (\Sigma \times N) \times M(F_i)$

Morphs that are phonologically the same but paired with different integers are assigned different semantic representations in the lexicon. They are instances of homonymous lexical entries. As before, the blocking rules are pairs of the first coordinates of lexical entries. However, keep in mind that now the first coordinates are morph-integer tuples.

When blocking co-exists with homonymous lexical entries, it is possible to specify multiple grammars for the same language. To demonstrate the types of grammars the general homonymy learner will induce from the data, consider the following two subsets of lexical entries for the overlapping affixes *-en* and *-t* in the German paradigm in table 3.5 in chapter 3.

(19) Two alternative grammars for German

Grammar 1 (LEX + BR)

(-en,1)	+group
(-t,1)	+part -speak, +group
(-t,2)	-part -speak, -group
BR: ((t,1),(en,1)); ((t,2),(en,1))	

Grammar 2 (LEX + BR)

(-t,1)	-speak
(-en,1)	+part +speak, +group
(-en,2)	-part -speak, +group

BR: ((en,1),(t,1)); ((en,1),(t,1))

In both of these grammars, one of the affixes is still the "elsewhere" form, while the other is split into two different lexical entries. Although there will sometimes be several possible "solutions" that a learner can find (i.e., several grammars for a single language), it will only converge on one of them depending on the order in which the input is presented. Namely, which morph will end up having an elsewhere status will depend on how early the learner sees it in comparison to the other morphs. This will become clearer when we consider the learning algorithm.

5.6.2 The algorithm

The General Homonymy Learner relies on the strategy of the Elsewhere learner which, as you recall, in turn relies on the strategy of the No-Homonymy learner. This nested relationship between the learners can be described as follows: we proceed under the assumption that the mapping between form and meaning is one-to-one until we have some positive evidence that this is not true, i.e., when invariant features of two different morphs become consistent. In this case we switch to the assumption that the discovered homonymy is due to an elsewhere-type of affix distribution. Recall that in such cases we can recover from overgeneralizations by introducing blocking rules rather than modifying lexical entries. However, if no blocking rule can be posited because we have evidence of an overlapping pattern (this is the new part), we switch to the least restrictive hypothesis

and consider the possibility that there is more than one homonymous lexical entry for the morph under consideration.

The strategy outlined above will succeed because at some finite point we can always determine whether we have overgeneralized either by (i) detecting homonymy, or (ii) detecting overlapping homonymy. The first kind of error is discovered as soon as the invariant features of two different morphs become consistent (see theorem 3). The second kind of error is discovered as soon as we have seen two competing morphs in the environments which are consistent with both of their invariant features or if we are about to add a contradictory blocking rule (see the definition of overlapping homonymy in (12)).

The General Homonymy Learner is different from the previous learners in that it has an additional component for detecting overlapping patterns and for postulating several homonymous lexical entries in the lexicon. The postulation of several lexical entries eliminates the overlaps and reduces the problem to the same scenario that the Elsewhere learner was facing, with one minor difference - now the input strings are ambiguous in the way they were not before (since a lexicon may contain more than one entry with the same phonological form).

Previously, we computed the invariant features of a morph incrementally by taking intersections of its current environment and the meaning of the morph in the current lexicon. Now, the current lexicon may contain several different lexical entries for a single phonological form, and so it is not immediately clear which of these we should choose to intersect with a given morph. When faced with such ambiguity, the learner will simply select the first in a list of possible lexical entries for the morph under consideration.[10] For instance, suppose there are two different morphemes in the current lexicon: $(-en, 1) - [1person, plural, fem.]$

[10] Alternatively, we could select the lexical entry whose current meaning is most similar to the environment we are considering. This could save us some time and make the lexicon shorter.

and $(-en, 2) - [3person, plural]$. When the learner encounters an input string containing $-en$, it will first proceed under the assumption that this morph is a realization of the morpheme $(en, 1)$. The learner will abandon this hypothesis only if it leads to a "dead end", i.e., if it is impossible to describe the distribution of the hypothesized morpheme with underspecification and blocking (in other words, if the distribution of $(en, 1)$ "overlaps" with the distribution of some other morph). In such a case, the learner will discard its current hypothesis without changing previous lexical entries and switch to the next possibility, namely the possibility that the morph $-en$ in the current input string is an instance of $(en, 2)$. If all possibilities lead to a dead end (i.e., an overlap), the algorithm will create a new entry for $-en$ assigning to it a new index and setting its second coordinate to the value of the environment in which $-en$ has just been seen.

In what follows, I step through this routine as it is presented in the Algorithm 'Main' (on page 146 in the pseudo-code format.

The first thing we do for each individual morph x is look up all morphemes in the current lexicon that have x as their first coordinate. I call this set of morphemes *homophones of x*. The *main* function has a while loop that goes through the homophones and calls another function *lexicalize* which is used to add a new hypothesized morpheme to the lexicon. If *lexicalize* returns *true* this means that an overlapping pattern was detected, in which case we continue in the while loop of the *main* function and try the next homophone. If $false$ is returned, then the new morpheme was successfully added to the lexicon and we break from the while-loop. If we have exhausted all homophones (i.e., *true* was returned on all of them), we create a new lexical item for x and add it to the lexicon. To do this, we need to compute a next integer to be paired with x (this is done in the function *nextIndex*). Next index of x will be set to 1 if x is not

currently in the lexicon, and to $k + 1$ where k is the largest integer currently associated with a morph x.

So, how does the algorithm detect overlapping homonymy? This is done in the function *lexicalize*. Recall that an overlapping pattern arises when some morph a occurs in the environment compatible with invariant features of another morph b, and vice versa: b occurs in an environment compatible with the invariant features of a. Additionally, overlaps can be detected by watching out for circular blocking rule chains.

If an overlapping pattern is detected, we immediately return *true*, which means that we abandon the current hypothesis and consider an alternative homophone. If no overlapping pattern has been detected, the algorithm proceeds to determine the blocking rules for all the competitors. If none of the competitors overlap with the current morpheme, all discovered blocking rules are added to BR.

For a better understanding of how this algorithm works, I go through a short example. Suppose that we have four universal binary features $f1, f2, f3, f4$, and that all strings in our language have length 1 (i.e., there is only one slot). Furthermore, features $f3$ and $f4$ are irrelevant, there are only two morphemes A and B, and they are in the following overlapping distribution:

(20) A hypothetical paradigm for language L

	+f1	$-f1$
+f2	A	B
$-f2$	B	A

Suppose that the text for this language begins as follows:

(21) A text for L

1. A $-f1, -f2, +f3, -f4$
2. A $-f1, -f2, -f3, -f4$
3. B $+f1, -f2, +f3, +f4$
4. A $+f1, +f2, +f3, -f4$
5. B $-f1, +f2, -f3, -f4$
6. B $-f1, +f2, +f3, -f4$
7. A $+f1, +f2, -f3, +f4$

After seeing the first three lines of text, our lexicon looks like this:

(22) LEX:

(A,1) $[-f1, -f2, -f4]$

(B,1) $[+f1, -f2, +f3, +f4]$

We are now processing the third input pair (A, $[+f1, +f2, +f3, -f4]$). First of all, we find all lexical entries in the current lexicon that have A as their first coordinate. There is only one of them. We call the function *lexicalize* and form a new lexical entry by taking an intersection of the current environment and the "old" meaning of A. As a result we get a potential morpheme $((A, 1)[-f4])$. Next, we check whether this new morpheme has any competitors in the current lexicon, i.e., other entries whose second coordinates are consistent with $[-f4]$. There are no such competitors (the condition of the while loop in *lexicalize* is not met because the length of the *competitors* set is 0), so we replace the old entry ((A,1),$[-f1, -f2, -f4]$) with the new entry $((A, 1)[-f4])$.

Now we move on to the next input pair (B,$[-f1, +f2, -f3, -f4]$). There is

only one entry in the lexicon for the morph B. We take an intersection with this entry and get a new potential morpheme $((B,1),\emptyset)$. This time, the set of competitors is not empty: $(A,1)$ is a competitor of $(B,1)$. Then, we check for an overlapping pattern: we look at all the expressions seen so far that are consistent with the current meanings of A and B (this is the set P in the Algorithm 4). The input pairs 1,2,4,and 5 will be in P because they are consistent with the feature set $\{\{-f4\} \cup \{\emptyset\}\}$. The condition in line 9 is met, i.e., P contains expressions whose first coordinates contain A and and those whose first coordinates contain B. Therefore, we know we found an overlapping pattern. This means we will not add $(B,\emptyset)$ to the lexicon, instead we set the variable *break* to true, and exit the while loop of *lexicalize*.

Because *lexicalize* returns true, we continue in the while-loop of the *main* function. But there is no other lexical entry with B as the first coordinate, so we form a new lexical entry for B with the index 2 and the semantic value $([-f1,+f2,-f3,-f4])$. We then try to lexicalize this new morpheme. This is now our second round in the *lexicalize* function. The new morpheme is still in competition with the lexical entry for $((A,1),[-f4])$. However, this time around no overlapping pattern is detected since P contains only one expression $B - ([-f1,+f2,-f3,-f4])$. From this we can infer that $(B,1)$ has to block $(A,1)$. So, after exiting the while-loop (with the variable *break* still set as $false$), we successfully add the new lexical entry and the new blocking rule to the grammar and exit the *lexicalize* function by returning $false$ and hence breaking from the while-loop in the *main* function. At this point, the lexicon looks as follows:

(23) LEX:
(A,1) $[-f4]$
(B,1) $[+f1,-f2,+f3,+f4]$

(B,2) $[-f1, +f2, -f3, -f4]$

BR: ((B,2),(A,1))

Next, we process the input pair 6. There are now two different lexical entries for B. We try to merge the current input with the first of these – $(B, 1)$. This leads to the potential morpheme $((B, 1), [+f3])$ which competes with $((A, 1), [-f4])$. We don't end up lexicalizing this potential morpheme because it is overlapping. Namely, we saw the feature set $[+f3, -f4]$ occur in the environments that were associated with both A and B. Therefore, we try the second entry - $(B, 2)$. Through intersection we get a potential morpheme $((B, 2), [-f1, +f2, -f4])$, which also competes with $(A, 1)$, but this competition can be resolved by a blocking that is already part of the grammar. So, we go ahead and replace the old entry for $(B, 2)$ with the new one, and now our lexicon looks like this:

(24) LEX:

(A,1) $[-f4]$

(B,1) $[+f1, -f2, +f3, +f4]$

(B,2) $[-f1, +f2, -f4]$

BR: ((B,2),(A,1))

Finally, consider the last input pair in our example. I leave it as an exercise for the reader to verify that the cross-situational intersection will lead to a new potential morpheme $(A, 1), \emptyset$, which will be successfully lexicalized (since no overlaps can be detected) with an addition of a new blocking rule $((B, 1), (A, 1))$.

It should be clear at this point that we are bound to converge on the correct grammar below, by continuing to rule out irrelevant features.

(25) LEX:

(A,1) $\emptyset$

(B,1) $[+f1, -f2]$

(B,2) $[-f1, +f2]$

BR: ((B,2),(A,1)); ((B,1),(A,1))

Notice that a somewhat different order of presentation of the input pairs could give us a different (but generatively equivalent) grammar:

(26) LEX:

(B,1) $\emptyset$

(A,1) $[+f1, +f2]$

(A,2) $[-f1, -f2]$

BR: ((A,2),(B,1)); ((A,1),(B,1))

Overall, this algorithm introduces homonymous lexical entries into the lexicon only when it is impossible to posit blocking rules to resolve lexical competition. By preventing the merging of morphs that would lead to an overlap, the learner essentially reduces the problem to the same situation that the elsewhere learner was faced with.

As the Elsewhere learner, the General Homonymy learner sometimes overgeneralizes. But eventually, it fixes overgeneralizations either by positing blocking rules or homonymous lexical entries. When no evidence has been seen to determine the direction of blocking or the presence of an overlapping pattern, the learner predicts free variation among several forms. These rare cases arise when the competitors under consideration are not currently overlapping and neither of them has been seen in any environment that is consistent with both of them.

In such cases, no blocking rules are added and the competition between lexical items is temporarily unresolved until the relevant data is encountered.

The General Homonymy learner will succeed because it correctly diagnoses when an overlapping pattern occurs (relying straight-forwardly on the definition of overlapping homonymy), and at that point it does not change any lexical items or introduce any blocking rules that are inconsistent with the data seen so far. Instead, it posits a new lexical entry for the morph in question.

In general, the algorithm is still relatively simple and consistent: at every point it either adds a new lexical item or a new blocking rule which correctly accounts for all the data seen so far. Since the algorithm is consistent and it operates in the finite space, we know that it is PAC-learnable (Anthony and Biggs, 1992, p.29).

Like the first elsewhere learner we considered, this algorithm also keeps a memory stack to which it can later refer. This stack is used for detection of overlaps. Undoubtedly, this algorithm can be made more efficient and perhaps also memoryless (assuming a fat text as before); however, I will not explore these option here.

Finally, it is also worth noting that the grammars this learner converges on are not necessarily most minimal grammars satisfying all the relevant requirements. This is particularly obvious for the languages with many overlaps. The order in which the input is processed will have a crucial impact on how the overlaps are resolved, and how many homonymous lexical entries are posited.

Algorithm 4 The General Homonymy Learner: main

Input: a text T for L consisting of pairs (s, e), where s has length p.
Output: a lexicon Lex $= Lex_1...Lex_p$, a set of blocking rules $BR = BR_1...BR_p$.

1: $Mem = \emptyset$
2: **for all** pairs (s, e) in T **do**
3: $Mem =$ add (s,e) to Mem
4: **for all** $s_i \in s$ **do**
5: $homophones \leftarrow \{((m, k), v) \in Lex_i | m = s_i\}$
6: $j \leftarrow 0$
7: fail $\leftarrow true$
8: **while** fail = true **do**
9: **if** $j \geq$ length of *homophones* **then**
10: index $\leftarrow$ **nextIndex**(s_i, Lex_i)
11: fail $\leftarrow$ **lexicalize**$(((s_i, index), e), ((*, 0), \emptyset), Lex_i)$ {((*,0),∅) is a dummy lexical entry}
12: **else**
13: $((m, k), v) \leftarrow homophones.(j)$
14: $newMeaning \leftarrow (v \bigcap e)$
15: fail $\leftarrow$ **lexicalize**$(((s_i, k), newMeaning), ((m, k), v), Lex_i)$
16: **end if**
17: j $\leftarrow j + 1$
18: **end while**
19: **end for**
20: **end for**

Algorithm 5 lexicalize

Input: new morpheme $((m,k),v)$, old morpheme $((m,k),oldMeaning)$, Lex_i, BR_i

Output: fail = false, if the new morpheme was successfully added to Lex_i, true otherwise

1: $competitors \leftarrow \{((m',k'),v') \in Lex_i | (m',k') \neq (m,k)$, & v' is consistent with $v\}$
2: $tempBR \leftarrow \emptyset$ {temprorary blocking rules}
3: $j \leftarrow 0$
4: break $\leftarrow false$
5: **while** $break = false$ and $j <$ the length of $competitors$ **do**
6: $((m',k'),v') \leftarrow competitors.(j)$
7: $j \leftarrow j+1$
8: $P \leftarrow \{(s,e) \in Mem \mid e$ is consistent with v' and $v\}$
9: **if** $\exists (s,e) \in P$, where $s_i = m$ **then**
10: **if** $\exists (s,e) \in P$, where $s_i = m'$ **then**
11: break $\leftarrow true$ {overlapping pattern}
12: **else**
13: **if** transitive closure of BR will have a contradiction with addition of $((m,k),(m',k'))$ to it **then**
14: break $\leftarrow true$ {overlapping pattern}
15: **else**
16: add $((m,k),(m',k'))$ to $tempBR$
17: **end if**
18: **end if**
19: **else**
20: **if** $\exists (s,e) \in P$, where $s_i = m'$ **then**
21: **if** transitive closure of BR will have a contradiction with addition of $((m',k'),(m,k))$ to it **then**
22: break $\leftarrow true$ {overlapping pattern}
23: **else**
24: add $((m',k'),(m,k))$ to $tempBR$
25: **end if**
26: **else**
27: {No evidence for overlaps or blocking rules.}
28: **end if**
29: **end if**
30: **end while**
31: **if** break = true **then**
32: **return** $true$
33: **else**
34: replace $((m,k),oldMeaning)$ with $((m,k),v)$ in Lex_i
35: $BR_i \leftarrow$ transitive closure of $BR_i \cup tempBr$
36: **return** $false$
37: **end if**

5.7 Discussion

5.7.1 Properties of the learners

In this chapter I presented three learners operating within increasingly larger and more complex learning spaces. The first learner was designed to learn languages with no-homonymy in the same slot. It followed a simple intersectional procedure to calculate invariant features and was equivalent to a learner that learns a set of monomials.

The second learner was a little more sophisticated: on top of calculating invariant features, it also added blocking rules to resolve competition among lexical items. As a result, this learner successfully learned languages with elsewhere homonymy but no overlapping distributions. In other words, it handled majority of inflectional paradigms, judging from the typological data on verbal inflection discussed in chapter 4.

Finally, the last (General Homonymy) learner had the additional power to add homonymous lexical entries when the competition among the morphs could not be resolved by any blocking rule, i.e., in presence of overlapping homonymy (free variation was explicitly ruled out).

The General Homonymy learner easily learns 1-1 form-meaning mappings using cross-situational intersections (in the same way as the No-Homonymy learner). It has to do more work in order to learn elsewhere distributions because those require checking for competitors and determining the appropriate blocking rules. This learner does it in a similar way as the Elsewhere learner, except it performs an additional check for overlapping patterns before positing blocking rules. When this check is positive, i.e., an overlap has been detected, the learner has to do still more work. In particular, when an overlap is discovered, the learner

abandons its current hypothesis and has to start from scratch, moving on to a different lexical item or forming a new one, calculating competitors, determining blocking, etc. In other words, in presence of overlaps, the algorithm goes through several passes of the function *lexicalize* (cf. the example in the previous section). Thus, this algorithm predicts that the overlapping homonymy requires more time and resources to learn. This prediction fits the empirical observation that such patterns are rare (and by hypothesis complex).

In short, the General Homonymy algorithm behaves in such a way that it is biased to rely on simple generalization strategies resulting in the learning of simple patterns (1-1 and elsewhere), but it has an ability to shift to more complex strategies when simple strategies are not sufficient. As discussed in the introduction to this chapter, the space of learning hypotheses can be thought of as structured into increasingly larger subsets with smaller (more restricted and simpler) subsets being preferred by the algorithm and therefore being empirically more probable.

If a language had an abundance of overlapping patterns, there would be little reason to propose a learner similar to the General Homonymy learner which relies on the idea of defaults. But because overlapping homonymy is rather rare (while elsewhere homonymy is common), defaults are still useful in describing and learning the grammars in a relatively simple fashion.

One other distinguishing property of my learner is that it overgeneralizes and subsequently corrects its overgeneralizations. The generalization method that the learner uses can be broken down into two parts. First, it generalizes in calculating invariant features which are used to find appropriate lexical specifications for the morphs. Since majority of morphs are not ambiguous, the invariant features alone are sufficient to define their meaning. But in cases of homonymy, invariant

features lead to overgeneralizations. Such overgeneralizations are later corrected either by blocking rules or by new lexical entries. The former method (positing blocking rules) is also generalizing. For example, if the learner has some evidence that a morph "a" blocks a morph "b" (without yet knowing precisely what the meaning of "a" or "b" is, and without seeing these morphs in all possible environments), it posits a blocking rule "$a \gg b$". This generalization method is safe due to the way the algorithm is structured. That is, we are guaranteed that no future data point will lead the algorithm to posit a contradictory blocking rule. This danger simply does not exist for the elsewhere homonyms (majority of the homonyms) because of the properties of the "elsewhere" patterns. As for the overlapping homonyms, this danger is eliminated by the algorithm since it checks for overlaps before positing the blocking rules.

How do invariant features help us in getting at the morphs' meanings? As I showed in this chapter, the invariant features are directly relevant to the meaning of non-ambiguous morphs. When it comes to the ambiguous morphs, the invariant features can be used as a first approximation or a rough cut that is made more precise with help of the blocking rules or the introduction of lexical homonymy.

In the rest of this discussion section, I consider predictions with regard to language acquisition and some remaining problems not addressed by the General Homonymy learner.

5.7.2 Predictions

Although the learner presented here rests on many idealizations, it already makes some interesting predictions with respect to the general trajectory of morphological acquisition.

One of the main predictions of this learner is the presence of overgeneralizations at intermediate learning stages and subsequent corrections of such overgeneralizations. The exact rate and types of overgeneralizations will depend on the order in which the learner is exposed to the input (which in turn could be connected to different frequencies of affixes).

In general, however, we would expect that, when the invariant features of two morphs are in a subset relationship and the more specific morph is less common, the more general morph would be temporarily used in the domain of the more specific morph. In cases in which the invariant features of different homophones are consistent but are overlapping or are in the subset relationship, we predict that these morphs should sometimes be used interchangeably, as though they were in free variation.

Some studies in language acquisition report several cases of overgeneralization errors in presence of homonymy (Ferdinand, 1996; Blom, 2003; Weerman et al., 2003; Berger-Morales, in progress), however they are difficult to evaluate with respect to the proposed learners for several reasons. First, there are many factors affecting children's output that are either irrelevant to learning or are not addressed by the current model (constraints on processing, phonological limitations, markedness, relative frequency of forms, etc.). Second, different language acquisition studies use different methodologies and different evaluation metrics for what counts as an overgeneralization, so that it is not immediately clear how to assess and compare such data. It is possible, that an artificial grammar learning experiment with children would be more beneficial for testing the above learning predictions.

We also predict that in presence of overlaps, it is possible that different children might arrive at slightly different grammars that nevertheless generate the

same input. The differences would manifest themselves in what forms are assumed to be defaults and what forms are listed as homonymous. Unfortunately this prediction is hard to test since we don't yet have a good way of determining which lexical representations in the mental lexicon have the default status.

Having such a method would also be useful in testing the psychological reality of the representations themselves (rather than the behavior of the learners). Yang (2005) considers a few possibilities for testing predictions of grammars that incorporate elsewhere statements of the type below:

IF $w = w_1$ **THEN** ...

ELSE IF $w = w_2$ **THEN** ...

...

ELSE IF $w = w_n$ **THEN** ...

ELSE apply R

More specifically, according to the "elsewhere condition serial search" model of processing that Yang assumes, the default rules should take longer to apply compared to other rules (provided that one controls for the stem-cluster frequency and the time course of rule application).

Lexical priming can also be potentially used to test the predictions of a model containing defaults. In particular, taking this model and the learner's output quite literally (i.e., assuming that the lexical entries derived by the learner correspond to the entries in the mental lexicon, while the blocking rules are in some separate grammatical component) we would predict that different instances of the elsewhere homonyms should prime each other, but instances of the overlapping homonyms should not prime each other. This is because on the standard interpretation, priming is due to re-activation of a recently accessed representation, and in the case of the elsewhere homonymy the learner I presented posits a

single underspecified lexical entry, while the same is not true for the overlapping homonymy.

5.7.3 Remaining problems

5.7.3.1 Free variation

When languages can have both free variation and overlapping homonymy, they are more challenging to learn. This is because once a learner discovers that a paradigm contains an overlapping distribution, it then still has to determine whether it is due to homonymy or to free variation. Free variation would be easy to spot if we already knew what the relevant features were. Then, by simply observing when the same sets of features are always associated with several different affixes, we could infer that such affixes stand in free variation. We could do this in the present scenario, but given a large number of universal features it might take a very long time before we discover that two different strings are always used in the same environments.

In general, the problem is this: in the initial stages of learning, free variation is hard to tell apart from overlapping homonymy. For instance, consider the following subset of input pairs:

(27) Hypothetical text

a $-f1-f2-f3$

a $-f1-f2+f3$

a $-f1+f2+f3$

b $-f1+f2-f3$

b $+f1+f2+f3$

Given this input we know that the paradigm contains an overlapping distribution, but it could either be due to free variation if $f3$ is irrelevant, or to overlapping homonymy if $f3$ is relevant. That is, the above input is consistent with a grammar in which $f3$ is irrelevant, and a is in free variation with b in some contexts:

(28) Grammar 1:
LEX: (a,1) - [-f1]; (b,1) - [+f2]

But it is also consistent with a grammar in which $f3$ is a relevant feature and there are two different entries for a:

(29) Grammar 2:
LEX: (a,1) - [-f1,-f2]; (a,2) - [-f1,+f2,+f3]; (b,1) - [+f2]
BR: ((a,2),(b,1))

It seems that to tackle both overlapping homonymy and free variation at the same time, it would be useful to exclude irrelevant features from consideration at some intermediate stage of learning. We know that a feature is irrelevant when it is not part of any lexical representation for any morphemes in a language, or when its value never affects the spell out of the strings. We could at some point guess what features are irrelevant by keeping track of those features that are always intersected out from every single morph in a given slot. Given this guess, we could evaluate whether the paradigm contains free variation by checking if any input pairs whose second coordinates have exactly the same *relevant* features contain different morphs in the same slot.

Alternatively, we could detect free variation whenever overlaps cannot be resolved by simply postulating more homonymous lexical entries. This would be

possible since overlaps that are due to free variation could never be removed. In other words, such overlaps are part of the grammar, they represent non-resolved competition that leads to use of several morphs interchangeably. I leave it to future research to work out exactly how this property of free variation could be used to detect it and how learning of free variation could be successfully incorporated into an on-line learning algorithm presented in the previous section.

5.7.3.2 Learning co-occurrence restrictions

Recall that one of our assumptions was that languages do not include co-occurrence restrictions conditioned by phonological or arbitrary lexical factors. We would like to eventually relax this assumption to be able to learn such co-occurrence restrictions which would form a basis for learning inflectional classes, as well as other types of non-semantic features like gender of inanimates, etc.

The simple learners that I have discussed, do not pay attention to the co-occurrence restrictions. But we can imagine making them more sophisticated so that they keep track of what subset of morphs a given morph can occur with. Once such subsets are correctly identified, the learner can attempt to find a conditioning factor for them. There is, of course, a whole set of new problems that such a learner would have to deal with if it attempts to generalize. For instance, if we generalize "bottom-up", when do we reach enough evidence to merge distinct subsets into a single set? Or, in other words, when do we conclude that morphs currently assigned to different subsets actually have the same distribution? If we generalize "top-down", assuming at first that all morphs can co-occur with all other morphs, when do we split them into the distinct subsets? And how do we recover from over- or under-generalizations that these strategies bring with them?

I am currently working on some ideas of how to solve this problem, but they are beyond the scope of this thesis.

CHAPTER 6

Summary

What I hope to have shown in this dissertation is that learning in the presence of homonymy is not a trivial task since languages contain unrestricted kinds of homonymy. However, the learner can still rely on a relatively simple learning strategy most of the time, taking advantage of the fact that majority of affix distributions are not overlapping, which is to say that the relationship between form and meaning is usually not arbitrary.

The learner presented here incorporates a bias for particular kinds of form-meaning mappings, namely those that can be described with underspecification and defaults. If this learner is on the right track, it suggests that the learning biases might be responsible for certain core properties of languages manifested in strong statistical tendencies. These tendencies organize the space of form-meaning patterns into a structured hierarchy, in which the higher is a pattern in the hierarchy, the more arbitrary and difficult it is for the speakers.

To show that inflectional form-meaning mappings are subject to strong statistical preferences, I examined two hypotheses about verbal agreement paradigms.

First, I considered the hypothesis that majority of inflectional paradigms do not include cases of true homonymy. This might strike one as surprising, since morphological descriptions of many languages include examples of what I called form identity, the scenario in which different paradigm cells are occupied by

the same phonological form. However, the empirical data on verbal agreement paradigms discussed in chapter 4 suggests that, in fact, agreement sub-paradigms containing homonymy amount to less than 20% of all paradigms.

Since homonymy is expected to be extremely frequent by pure chance (see chapter 4), the fact that it is rather limited cannot be accidental. The natural and widely accepted explanation for why languages avoid homonymy is that it poses a problem for processing and learning. For instance, as I showed in the course of describing the learning algorithms, presence of homonymy leads to overgeneralization errors which require additional effort to eliminate. From the processing point of view, presence of homonymy requires additional resources for sense disambiguation.[1]

The second constraint on homonymy I explored is more controversial. It has to do with preference of homonymy patterns that can be described with defaults compared to those that cannot (i.e., the overlapping homonymy). Overlapping homonymy involves several morphs whose distribution is intertwined in a complex way, so it appears arbitrary. We saw that overlapping patterns are expected to be quite frequent by chance and that, contrary to this expectation, they are very rare in the verbal agreement paradigms.

Interestingly, similarly to the absence of homonymy, the scarcity of overlapping patterns is also advantageous for learning. In particular, as the Elsewhere learner shows, excluding overlaps from the hypothesis space allows one to rely on a simple procedure for correcting overgeneralizations due to homonymy. This procedure requires no backtracking or changing of the previous hypothesis. It

[1]However, the view that homonymy is problematic in this way is not incompatible with the view that it might also be beneficial when we change our perspective. Namely, it is sometimes suggested that homonymy helps to reduce the number of distinct phonological units. As far as I know, there is no strong evidence for this claim, however, it it turns out to be true, then we could modify the current learning algorithm to model divergent pressures posed by homonymy.

merely consists of adding blocking rules that in effect restrict the distribution of the elsewhere homophones.

In the presence of overlaps, the same strategy can be largely maintained with an additional complication of adding homonymous lexical entries to resolve ambiguity of overlapping homophones. In principle, this complication could lead to very inelegant and convoluted grammars, but since overlapping homonymy is quite rare, it is still possible to keep both the learning procedure and the resulting grammars simple.

Thus, the learning model proposed here fits the statistical findings: it requires most complex calculations for learning overlapping patterns, and least complex calculations for learning non-homonymous mappings with elsewhere homonymy falling in between. At the same time, this learning strategy is based on a natural and intuitive generalization method which lends credibility to the idea that people rely on non-monotonic reasoning. Recall that, in somewhat simplified terms, the type of non-monotonic learning we employ here can be thought of as holding on to general rules and memorizing exceptions to it as long it is possible to do so. I expect that psycholinguistic evidence would shed more light on whether this is indeed the kind of strategy people use in learning meanings of morphemes.

Bibliography

Adger, David. 2006. Combinatorial Variation. *Journal of Linguistics* 42:503–530.

Albro, Daniel. 1997. A morpho-phonological learner. URL `www.linguistics.ucla.edu/people/grads/albro/mlearner.pdf`, ms., UCLA.

Anderson, Stephen R. 1992. *A-Morphous Morphology*. Cambridge University Press, Cambridge.

Anthony, M. and N. Biggs. 1992. *Computational Learning Theory*. Cambridge University Press, Cambridge.

Antoniou, Grigoris. 1997. *Nonmonotonic Reasoning*. MIT Press, Cambridge, Massachusetts, London, England.

Aronoff, M. 1976. *Word Formation in Generative Grammar*. MIT Press, Cabrbidge, Mass.

Aslin, R.N., J.R.Saffran, and E.L.Newport. 1998. Computation of conditional probability statistics by 8-month old infants. *Psychological Science* 9:321–324.

Baerman, Matthew. 2004. Typology and the formal modeling of syncretism. In *Yearbook of Morphology*, edited by G. Booij and J. van Marle. Kluwer, pages 41–72.

Baerman, Matthew, Dunstan Brown, and Greville G. Corbett. 2005. *The Syntax Morphology Interface: A Study of Syncretism*. Cambridge University Press, New York.

Baroni, Marco. 2003. Distribution-driven Morpheme Discovery: A computational/experimental study. In *Yearbook of Morphology*, edited by Geert Booj and Jaap van Marle. Springer, Dortrecht.

Bazell, C.E. 1960. A question of syncretism and analogy. *Transactions of the Philological Society* :1–12.

Bentin, S. and L. Feldman. 1990. The contribution of morphological and semantic relatedness to repetition priming at short and long lags: Evidence from Hebrew. *The Quarterly Journal of Experimental Psychology* 42:693–711.

Benveniste, Emile. 1971. *The nature of pronouns. Problems in general linguistics*. University of Miami Press, Coral Gables, FL.

Berger-Morales, Julia. in progress. The Acquisition of Inflectional Morphology in the Nominal Domain of German. Ph.D. thesis, UCLA.

Bierwisch, Manfred. 2006. Luxury in Natural Language. url: http://www.zas.gwz-berlin.de/publications/40-60-puzzles-for-krifka/pdf/bierwisch.pdf.

Blom, E. 2003. From Root Infinitive to Finite Sentence. Ph.D. thesis, Utrecht University.

Bloom, Paul. 2000. *How Children Learn the Meanings of Words*. MIT Press, Cambridge, MA.

Blum, L. and M. Blum. 1975. Toward a mathematical theory of inductive inference. *Information and Control* 28:125–155.

Booth, A.E. and S.R. Waxman. 2002. Word Learning is "Smart": Evidence

that Conceptual Information Affects Preschoolers Extension of Novel Words. *Cognition* 87(3):215–218.

Brent, M. 1999. An Efficient Probabilistically Sound Algorithm for Segmentation and Word Discovery. *Machine Learning* 34:71–106.

Butterworth, B. 1983. Lexical Representation. In *Developement, writing, and other language processes*, edited by B. Butterworth, 2. Academic Press, London.

Bybee, J. L. 1985. *Morphology: A study of the relation between meaning and form*. Benjamins, Philadelphia.

Cambell, T. and C. Yang. 2005. Mechanisms and Constraints in Word Segmentation. Manuscript, Yale University.

Caramazza, A., A. Laudanna, and C. Romani. 1988. Lexical access and inflectional morphology. *Cognition* 28:297–332.

Carlson, Lauri. 2005. Inducing a morphological transducer from inflectional paradigms. In *Inquiries into Words, Constraints and Contexts*, edited by Ann Copestake, CSLI Studies in Computational Linguistics ONLINE. CSLI Publications, http://cslipublications.stanford.edu/site/SCLO.html.

Carstairs, Andrew. 1984. Outlines of a constraint on syncretism. *Folia Linguistica* 18:73–85.

Carstairs, Andrew. 1998. How lexical semantics constrains inflectional allomorphy. In *Yearbook of Morphology 1997*, edited by Geert Booij and Jaap van Marle. Kluwer, Dortrecht, Boston, London, pages 1–24.

Chomsky, Noam and Morris Halle. 1968. *The sound pattern of English*. Harper & Row, New York.

Clahsen, H. 1999. Lexical entries and rules of language: A multidisciplinary study of German inflection. *Behavioral and Brain Sciences* 22:991–1060.

Corbett, Greville. 2000. *Number*. Cambridge University Press, Cambridge.

Cysouw, Michael. 2001. The paradigmatic structure of person marking. Ph.D. thesis, University of Nijmegen.

de Marcken, C. 1996. Unsupervised Language Acquisition. Ph.D. thesis, MIT.

Dhagat, Aditi and Lisa Hellerstein. 1994. PAC learning with irrelevant attributes. In *Proceedings of the 35rd Annual Symposium on Foundations of Computer Science*. IEEE Computer Society Press, Los Alamitos, CA, pages 64–74. URL `citeseer.ist.psu.edu/dhagat94pac.html`.

Ferdinand, A. 1996. *The Development of Functional Categories: The Acquisition of the Subject in French*. Holland Academic Graphics, The Hague.

Finch, S. and N. Chater. 1992. Bootstrapping syntactic categories using statistical methods. In *Machine Learning of Natural Language*, edited by D. Daelemans and W. Powers. Tilburg, NL, pages 229–236.

Fisher, Cynthia, G. Hall, S. Rakowitz, and L. Gleitman. 1994. When is it better to recieve than to give: syntactic and conceptual constraints on vocabulary growth. *Lingua* 92:333–375.

Fodor, Jerry A. 1998. *Concepts; Where cognitive science went wrong*. The 1996 John Locke Lectures. Oxford University Press.

Fodor, Jerry A., M. Garret, E. Walker, and C. Parkes. 1980. Against definitions. *Cognition* 8.

Forchheimer, Paul. 1953. *The category of person in language*. Walter de Gruyter, Berlin.

Frauenfelder, U.H. and R. Schreuder. 1992. Constraining psycholinguistic models of morphological processing and representation: The role of productivity. In *Yearbook of Morphology 1991.*, edited by G.E. Booij and J. van Marle. Kluwer, Dordrecht.

Gleitman, Lila. 1990. The Structural sources of verb meaning. *Language Acquisition* 1:3–55.

Gold, E. Mark. 1967. Language identification in the limit. *Information and Control* 10(5).

Goldfarb, L. 1986. Metric Data Models and Associative Memories. In *Proceedings of the 8th IASTED International Symposium on Robotics and Artificial Intelligence: Identification and Pattern Recognition*, volume 3. Toulouse, France, pages 53–73.

Goldsmith, John. 2001. Unsupervised Learning of a Morphology of a Natural Language. *Computational Linguistics* 27:153–198.

Gonnerman, Laura. 1999. Morphology and the lexicon: Exploring the semantics-phonology interface. Ph.D. thesis, University of Southern California.

Greenberg, Joseph. 1963. *Universals of language*. MIT Press, Cambridge.

Halle, Morris. 1997. Distributed Morphology: Impoverishment and Fission. *MIT Working Papers in Linguistics* 30:425–449.

Halle, Morris and Alec Marantz. 1993. Distributed Morphology and the Pieces of Inflection. In *The View from Building 20*, edited by K. Hale and S. J. Keyser. MIT Press, Cambridge, Mass., pages 111–176.

Halle, Morris and Alec Marantz. 1994. Some key features of Distributed Morphology. In *MITWPL 21: Papers on phonology and morphology*, edited by Andrew Carnie and Heidi Harley. MITWPL, Cambridge, pages 275–288.

Hardman, M.J. 2001. *Aymara*. Studies in Native American Linguistics 35. Lincom.

Harley, Heidi and Elizabeth Ritter. 2002. Person and number in pronouns: A feature-geometric analysis. *Language* 78:482–526.

Haspelmath, M. 2002. *Understanding Morphology*. Language Series. Arnold, London.

Haspelmath, M., M. S. Dryer, D. Gil, and B. Comrie, eds. 2005. *The world atlas of language structures*. Oxford University Press, London.

Hayes, B. 2005. Class notes on Morphology, UCLA.

Heinz, Jeff. 2007. Inductive Learning of Phonotactic Patterns. Ph.D. thesis, UCLA.

Hjelmslev, Louis. 1935. *La catégorie des cas*. Universitetsforlaget, Aarhus.

Hodges, Wilfrid. 2000. A Context Principle. In *Proceedings of Munich Conference on Intentionality*, edited by Reinhart Kahle.

Ide, Nancy and Jean Veronis. 1998. Word Sense disambiguation: the State of the Art. *Computational Linguistics* 24.

Imai, M. and D. Gentner. 1999. A Cross-linguistics Study of Early Word Meaning: Universal Ontology and Linguistic Influence. In *The Emergence of Language. Carnegie Mellon Symposium on Cognition*, edited by B. MacWhinney. Lawrence Erlbaum, Mahwah, N.J.

Jakobson, R. 1939. Signe zéro. Mélanges de linguistique offerts á Charles Bally. Reprinted in "Russian and Slavic Grammar", Mouton Publishers, 1984.

Jakobson, Roman. 1936. Beitrag zur allgemeinen Kasuslehre: Gesamtbedeutungen der russischen Kasus. *Travaux du Cercle Linguistique de Prague* 6:240–288.

Jakobson, Roman. 1971. *Selected writings*, chapter Shifters, verbal categories, and the Russian verb. Word and language. Mouton, The Hague, pages 130–147.

Johnson, E.K. and P.W. Jusczyk. 2001. Word Segmentation by 8-month-olds: When Speech Cues Count More than Statistics. *Journal of Memory and Language* 44:548–567.

Jones, S. and L. Smith. 2002. How Children Know the Relevant Properties for Generalizing Object Names. *Developemental Science* 5:219–232.

Ke, J. 2004. Self-organization and Language Evolution: System, Population and Individual. Ph.D. thesis, City University of Hong Kong.

Kiparsky, Paul. 1973. Elsewhere in Phonology. In *Festshrift for Morris Halle*, edited by P. Kiparsky and S. Anderson. Holt, Rinehart and Winston, New York.

Kobele, G.M., J. Riggle, T. Collier, Y. Lee, Y. Lin, Yao, C. Taylor, and E. Stabler. 2003. Grounding as learning. In *Language*

Evolution and Computation Workshop. ESSLLI, pages 201–225. URL `http://taylor0.biology.ucla.edu/al/`.

Kurimo, Mikko, Puurula, Arisoy, Siivola, Hirsimäki, Pylkkönen, Alumäe, and Saraclar. 2006. Unlimited vocabulary speech recognition for agglutinative langauges. In *Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*. New York, pages 487–494.

Lieber, Rochelle. 1992. *Deconstructing Morphology*. University of Chicago Press, Chicago.

Luraghi, Silvia. 1987. Patterns of case syncretism in Indo-European languages. In *Papers from the Seventh International Conference on Historical Linguistics*. John Benjamins, Amsterdam, pages 355–375.

Luraghi, Silvia. 2000. Synkretismus. In *Morphologie: Ein Handbuch zur Flexion und Wortbildung*, edited by Geerd Booij, Christian Lehmann, and Joachim Mugdan. Mouton de Gruyter, Berlin.

Marantz, Alec. 1997. No escape from syntax: Dont try morphological analysis in the privacy of your own lexicon. In *21st Annual Penn Linguistics Colloquium*. University of Pennsylvania, pages 201–225.

Marchman, Virginia, Kim Plunkett, and Judith Goodman. 1997. Overregularization in English plural and past tense. *Child Language* 24:767–779.

Marcus, Gary, Steven Pinker, Michael Ullman, Michelle Hollander, T. John Rosen, and Fei Xu. 1992. Overregularization in language acquisition. *Monographs of the Society for Research in Child Development* 57(4). Includes commentary by Harold Clahsen.

Marcus, Gary F. 1995. The acquisition of the English past tense in children and multilayered connectionist networks. *Cognition* 56:271–279.

Markman, E.S. 1984. The Acquisition and Hierarchichal Organization of Categories by Children. In *Origins of Cognitive Skills*, edited by C. Sophian. Lawrence Erlbaum, Hillsdale, N.J.

Markman, E.S. 1989. *Categorization and Naming in Children*. MIT Press, Cambridge, MA.

Mattys, S.L., P.W. Jusczyk, P.A. Luce, and J.L. Morgan. 1999. Phonotactic and Prosodic Effects on Word Segmentation in Infants. *Cognitive Psychology* 38:465–494.

Meiser, G. 1993. Syncretism in Indo-European languages. *Transactions of the Philological Society* 90:187–218.

Mintz, T.H. 2002. Category Induction from Distributional Cues in an Artificial Language. *Memory and Cognition* 30:678–686.

Muller, Gereon. 2004. On Decomposing Inflection Class Features: Syncretism in Russian Noun Inflection. In *Explorations in Nominal Inflection*, edited by Gereon Muller, Lutz Gunkel, and Gisela Zifonun. Mouton de Gruyter, Berlin.

Murane, E. 1974. *Daga Grammar. From morpheme to discourse.*. Norman, Oklahoma:SIL.

Noyer, Rolf. 1998. Impoverishment theory and morphosyntactic markedness. In *Morphology and its Relation to Phonology and Syntax*, edited by et al Lapointe. CSLI, Stanford, pages 264–285.

Osherson, Daniel, Scott Weinstein, and Michael Stob. 1986. *Systems that learn*. MIT Press, Cambridge, Massachusettes.

Pinker, S. 1991. Rules of Language. *Science* 253:530–534.

Pinker, Stephen. 1989. *Learnability and Cognition*. The MIT Press, Cambridge,MA.

Pitt, David. 1999. In defense of definitions. *Philological Psychology* 12(2):139–159.

Plank, Frans. 1980. Encoding Grammatical Relations: Acceptable and Unacceptable Non-distinctness. In *Historical Morphology*, edited by Jacek Fisiak. Mouton, The Hague.

Plank, Frans. 1986. Paradigm size, morphology typology, and universal economy. *Folia Linguistica* 20:29–48.

Prince, Alan and Paul Smolensky. 1993. Optimality Theory: Constraint interaction in Generative Grammar. Rutgers University Center for Cognitive Science Technical Report 2.

Regier, Terry. 2003. Emergent constraints on word-learning: A computational review. *Trends in Cognitive Sciences* 7:263–268.

Reh, M. 1985. *Die Krongo-Sprache*. Dietrich Reimer, Berlin.

Rissanen, J. 1978. Modeling By Shortest Data Description. *Automatica* 14:465–471.

Rota, G.C. 1964. On the foundations of combinatorial theory. *Zeitschrift Fur Wahrscheinlichkeitstheorie Und Verwandte Gebietee* 2:340–368.

Saffran, J.R., R.N. Aslin, and E.L. Newport. 1996. Statistical Learning by 8-month Old Infants. *Science* 274:1926–1928.

Santelmann, L., P. Jusczyk, and M. Huber. 2003. Infants Attention to Affixes. In *Jusczyk Lab Final Report*, edited by D.Houston, A.Seidl, G.Hollich, E.Johnson, and A.Jusczyk. http://hincapie.psych.purdue.edu/Jusczyk.

Sauerland, Uli. 1995. The Late Insertion of Germanic Inflection. Ms.

Seidenberg, M.S. and J.L. McClelland. 1989. A Distributed, Developmental Model of Word Recognition and Naming. *Psychological Review* 96:523–568.

Siewierska, A. 2004. *Person*. Cambridge University Press, New York.

Siskind, Jeffrey Mark. 1996. A computational study of cross-situational techniques for learning word-to-meaning mappings. *Cognition* 61(1-2):1–38.

Skinner, B.F. 1957. *Verbal Behavior*. Appleton-Century-Crofts, New York.

Smith, A.D.M. 2003. Intelligent Meaning Creation in a Clumpy World Helps Communication. *Artificial Life* 9(2):559–574.

Smith, L.B., S. Jones, L. Gershkoff-Stowe, and S. Samuelson. 1999. Toward a Universal Law of Generalization for Psychological Science. *Science* 237:1317–1323.

Stabler, E. forthcoming. Computational models of language universals: expressiveness, learnability and consequences. In *Language Universals*, edited by Morten Christiansen, Chris Collins, and Shimon Edelman. Oxford University Press.

Stockall, L. and A. Marantz. 2006. A single route, full decomposition model of morphological complexity: MEG evidence. *The Mental Lexicon* 1(1):85–123.

Strauss, Sidney and Ruth Stavy. 1982. *U-shaped behaiviral growth*. Developmental Psychology. Academic Press, New York.

Stump, Gregory. 1993. On rules of referral. *Language* 69:449–479.

Stump, Gregory T. 2001. *Inflectional Morphology: A Theory of Paradigm Structure*. Cambridge University Press, Cambridge.

Sumbatova, Nina and Rasul Mutalov. 2003. *Grammar of Icari Dargwa*. Languages of the World Mateirals 92. Lincom Europa, Muenchen.

Thiessen, Erik and Jenny Saffran. 2007. Learning to Learn: Infants' Acquisition of Stress-Based Strategies for Word Segmentation. *Language Learning and Development* 3(1):73–100.

Thompson, C.A. and R.J. Mooney. 2003. Acquiring Word-Meaning Mappings for Natural Language Interfaces. *Journal of Artificial Intelligence* 18:1–44.

Trommer, J. 2003. The interaction of morphology and syntax in affix order. In *Yearbook of Morphology*, edited by G. Booij and J. van der Marle. Kluwer, Dordrecht.

Trueswell, J.C., I. Sekerina, N.M. Hill, and M.L. Logrip. 1993. The kindergarten-path effect: studying on-line sentence processing in young children. *Cognition* 73:89–134.

Valiant, Leslie. 1984a. A theory of the learnable. *Communications of the Association for Computing Machinery* 27:1134–1142.

Valiant, L.G. 1984b. A Theory of the Learnable. *CACM* 17(11):1134–1142.

VanWagenen, Sarah. 2005. The Morphologically Organized Mental Lexicon: Further Experimental Evidence. Master's thesis, University of California, Los Angeles.

Vapnik, Vladimir. 2000. *The Nature of Statistical Learning Theory*. Springer, New-York, 2 edition.

Vogt, Paul. 2003. Investigating social interaction strategies for bootstrapping lexicon development. *Journal of Artificial Societies and Social Simulation* 6(1).

Wallace, C.S. and D.M. Boulton. 1968. An information measure for classification. *Computer Journal* 11:185–194.

Weerman, F., J. Bischop, and L. Punt. 2003. L1 and L2 acquisition of Dutch adjectival inflection. Paper presented at Generative Approaches to Language Acquisition, Utrecht, the Netherlands.

Werker, J.F. 1989. Becoming a Native Listener: A Developmental Perspective on Human Speech Perception. *American Scientist* 77(1):54–59.

Williams, Edwin. 1994. Remarks on lexical knowledge. *Lingua* 97:7–34.

Williams, Edwin and Anna-Maria DiScullio. 1987. *On the definition of a word*. MIT Press, Cambridge MA.

Wittgenstein, L. 1953. *Philosophical Investigations*. Blackwell, Oxford.

Wunderlich, Dieter. 2004. Is there any need for the concept of directional syncretism? In *Explorations in Nominal Inflection*. Mouton de Gruyter, Berlin, pages 373–395.

Yang, Charles. 2005. On productivity. *Yearbook of Language Variation* 5:333–370.

Zwicky, Arnold. 1985. How to describe inflection. *BLS* 11:371–386.