Minimal Generative Explanations: A Middle Ground between Neurons and Triggers

Michael Brent
Department of Cognitive Science
Johns Hopkins University
Baltimore, MD 21218
michael@mail.cog.jhu.edu

Abstract

This paper describes a class of procedures for discovering linguistic structure, along with some specific procedures and measures of their effectiveness. This approach is well-suited to problems like learning the forms of words from connected speech, learning word formation rules, and learning phonotactic constraints and phonological rules. These procedures acquire a symbolic representation, such as a list of word forms, a list of morphemes, or a set of context sensitive rules, each of which serves as the language-particular component of a generative grammar. Each procedure considers only a clearly defined set of possible generative grammars. This hypothesis space can be thought of as the procedure's "universal grammar". Procedures are evaluated for effectiveness by computer simulation on input consisting of naturally occurring language. Thus, they must be robust. That is, small changes to the input must lead to little or no change in the conclusions. This research program resembles the connectionist program in its focus on phenomena like word-segmentation, morphology, and phonology, its emphasis on robustness, and its reliance on computer simulation. However, it is closer to parameter setting and learnability theory in its focus on learning generative grammars selected from a clearly defined hypothesis space, or "universal grammar". Further, to the extent that connectionism is about neural implementations while parameter setting and learnability theory are about universal grammars, the study of effective procedures for language acquisition stands at an intermediate level of abstraction.

1 Introduction

Recent advances in the theory of induction have made it possible to design algorithms that discover generative grammars for some linguistic regularities. This paper explains the principle behind these algorithms and presents one particular algorithm for the discovery of morphemic suffixes. Implementing such algorithms and running them on naturally occurring language can shed light on the informational structure of natural languages, the robustness of language acquisition, and the value of search heuristics for language acquisition.

Informational Structure. Consider the various domains of linguistic knowledge, such as the forms of words, their meanings, their syntactic properties, the forms of morphemes, the morphotactic rules governing their combination, phonotactic constraints, and so on. To what extent is knowledge in each domain useful or necessary for learning about each of the other domains?

Robustness. Language acquisition is robust in the sense that garbled or ungrammatical input does not have catastrophic effects on the outcome. In fact, at naturally occurring frequencies it seems to have no effect at all. Further, creolization processes suggest that children are attracted to linguistic rules and tend to discount inconsistent evidence.[1] What do these properties have in common, and what kinds of language learning procedures have them?

Controlling Search. Even when known universal constraints are taken into account, the number of possible hypotheses consistent with a given linguistic input is generally so large that evaluating them all is computationally intractible. For example, the phonological effects of morphological processes in the world's languages are extremely diverse, including suffixes, prefixes, infixes, circumfixes, ablaut/umlaut, vowel-tier morphemes, tonal morphemes, metatheses, and truncations (Anderson, 1992). One technique that might aid children in the identification of morphemes is search ordering, where the most likely hypotheses are explored first. For example, suffixation appears to be the most common effect of morphological processes in the world's languages, and all languages in Greenberg's survey that have non-affixal morphology also have prefixes, suffixes, or both (Greenberg, 1966). Thus, it would make sense for children to look for suffixes and pre-

[1] Newport (1993) reports on a deaf child raised by deaf parents who are late learners of American Sign Language. Although the parents use a particular morpheme in only 65% of obligatory contexts, the child regularizes it, reaching about 90% of obligatory contexts.

fixes before looking for metatheses and truncations. Search ordering would clearly speed the identification of suffixes, as compared to searching for all sorts of phonological effects at once, and since suffixes are so common, the average rate of acquisition would be improved too. But looking for suffixes first might actually speed the acquisition of non-affixal morphology as well. The rapid discovery of some suffixes might provide the child with a toe-hold on the language's morphology, making possible partial analysis of the input and thereby simplifying the search for other morphemes. How much can search ordering help in language acquisition? What other search heuristics exist, and how much do they help?

These questions can be investigated by devising algorithms for learning in a particular domain, such as morpheme discovery, implementing them as computer programs, and running the programs on naturally occurring language. The informational structure of languages can be investigated by varying the information provided from domains other than the one to be learned. For example, how much does knowing the major syntactic category of each word help in discovering morphemic suffixes? This question is addressed in Section 3. If experiments on naturally occurring input leave any doubt about the robustness of an algorithm, it can be tested on input that has been artificially corrupted in various ways and to various degrees. Finally, computer simulation can be used to explore the effects of various constraints on the hypothesis space and heuristics for searching it, both in terms of the time required for learning and the outcome. For example, given an effective procedure for identifying morphemic suffixes, what happens when infixes and circumfixes are considered too? How much longer does it take to find the suffixes, to what extent are incidental regularities mistaken for infixes and circumfixes, and how does that affect which suffixes are found?

This paper presents a robust algorithm for discovering morphemic suffixes. This algorithm is not linguistically universal, in the sense that there are morphological processes other than suffixation. Further, it is not intended as a detailed model of how children process each input utterance. Rather, it should be seen as a tool for evaluating certain strategies that may be used by children's language acquisition algorithms. Ultimately, this algorithm may lead to others that are more linguistically universal and more faithful to the incremental nature of language acquisition, but that awaits further research. Finally, as a practical matter, the experiments that have been done to date use as input journalistic text in orthographic form. The conditions of child language acquisition would be more accurately reflected by phonetic transcripts of child-directed speech, and such experiments are planned. Nonetheless, the types of morphological phenomena to be found in the two forms of input are, at an appropriate level of abstraction, similar.

2 Generative Explanations

The algorithms presented here are based on an idea that Rissanen (1983) dubbed the Minimum Description Length Principle. This idea has deep roots in the theory of probability and computation, and, ultimately, in the philosophy of science (Li and Vitányi, 1992). Recently, practical inference procedures using this idea have been proposed by researchers in statistics (Rissanen, 1983; Rissanen, 1986) and artificial intelligence (Quinlan and Rivest, 1989; Ellison, 1991; Ellison, 1992).

Language acquisition procedures based on the Minimum Description Length Principle are perhaps better described as Minimum Generative Explanation procedures. These procedures attempt to find a generative explanation for the regularities in the input. Part of a generative explanation is a *theory* about how the input is generated. A generative theory in this sense could be something as simple as the theory that utterances are generated by selecting words from a lexicon and concatenating them. An *instantiation* of this theory would include a hypothesis about the word forms in the lexicon from which the input was generated. Such an instantiation would provide a partial explanation for the fact that *the* and *dog* occur repeatedly in the input in various different contexts, while *thed* and *og* appear in a very limited range of contexts — frequently together and in that order. In this way, identifying the word stock of the input language from connected speech can be cast as finding a language-particular word list for the concatenation theory, which itself is part of a search for a generative explanation of regularities in the input.

General theories such as the one that sentences are generated by concatenating words should be thought as heuristics for learning — they can be useful without explaining all the regularities in the input. For example, the word concatenation theory leaves many regularities unexplained, including those created by phonotactic, morphological, syntactic, and semantic constraints. It is possible that those regularities might overwhelm and confuse the regularities due to word forms, making it impossible to discover word forms using phonological information alone. It is also possible that, for example, phonotactic, morphological, and lexical regularities in the sound patterns of language can be detected independently of one another — in effect, that they constitute distinct signals on the same channel. This is the type of question about the informational structure of language which I hope to answer.

Another simple generative theory is that English words are formed by concatenating a stem and a suffix, each chosen from a fixed set. This theory is somewhat less accurate than the theory about sentences being generated by concatenating words — for instance, more than one suffix may be used in forming a given word. Further, phonological and orthographic rules often adjust a concatenated stem and suffix, as in *watches* (not "watchs") and *boxes* (not "boxs"). Nonetheless,

the concatenative morphology theory is not a bad zeroth-order approximation to English word formation. A somewhat refined theory is that, after concatenation of stem and suffix, phonological segments (or letters) are inserted or deleted at the morpheme boundary, in a way that depends on the original segments surrounding the boundary. Another, independent refinement is the theory that, once a stem and a suffix have been chosen, a syntactic category is chosen from among those available for the suffix. Instantiations of these theories, including specific stems, suffixes, adjustment rules, and so on, would help the learner explain why the observed input contains the patterns it does.

Minimum Generative Explanation procedures attempt to find the best instantiation of the general theory, given a certain body of input, where best means most accurate at predicting which utterances are acceptable and which are not. This problem can be separated into two subproblems. First, given two instantiations, how can the learner decide which one is better, in view of the input? Second, how can the learner generate some plausible instantiations to evaluate and compare?

Given a general theory, a *generative explanation* of the input consists of an instantiated theory plus an accounting of the properties of the input that are *not* explained by the theory. For example, the instantiated theory might include the assertions that *dog* is a possible stem and *-s* is a possible suffix, but it could not predict whether or not the word *dogs* will appear in some particular input sample — that depends on what people choose to talk about, which is outside the domain of the theory of word formation. An explanation for the occurrence of the word *dogs* in a given sample, then, has two parts: first, the theory that input words are generated by a procedure that is capable of generating *dogs* because *dog* is on its stem list and *-s* is on its suffix list; second, the assertion that the stem *dog* and the suffix *-s* happen to have been chosen for concatenation at some particular moment, due to unknown factors. Each of these two parts constitutes a stipulation. Once accepted, these two stipulations completely explain the input, including its predictable and its unpredictable properties. Minimum Generative Explanation algorithms formalize the venerable notion that the shortest explanation is most likely to predict future observations.

For the sake of illustration, consider the input sample consisting of the words shown in the left hand table of Figure 1. The remaining tables of Figure 1 provide a plausible generative explanation for the presence of these words in the input. Namely, words in the input are generated by a procedure that concatenates a stem from the stem table with a suffix from the suffix table, and the stems and suffixes that happen to have been concatenated to form the input sample are as indicated on the word table. The word table is represented here as a sequence of pairs of indices, or code words. The stem table shows the correspondence between stems and stem code words, while the suffix table shows the same for suffix code words. In order to allow each word to be represented uniformly as a stem and a suffix, each monomorphemic word is assigned a "null suffix," denoted by the symbol ϵ.

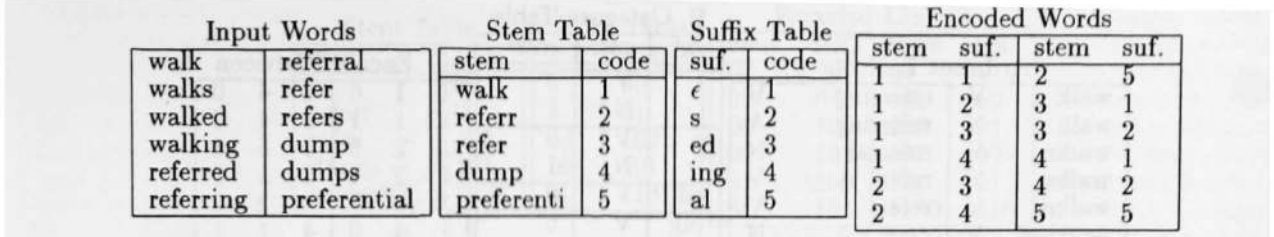
Now consider two alternative explanations that rely on different suffixes and stems. In Figure 2, the first row of tables shows a generative explanation in which each suffix is one character longer than before, while the second row shows an explanation in which each is one character shorter. Each of the three explanations is equally consistent with the data, but the first one seems more likely to accurately predict which words are acceptable than the latter two. For example, the explanation using long suffixes predicts that *dumring* and *dumked* are possible English words, whereas *dumped* and *dumping* are not. The first analysis correctly predicts the opposite, even though *dumped* and *dumping* have never been observed. (Of course, even first analysis makes some faulty predictions, since the general theory it instantiates ignores many of the constraints on English word formation.)

The difference in the predictive accuracy of these three theories can be understood in terms of the widely held intuition that the shortest explanation is the most likely one. Throughout the modern period, scientists have used this intuition, which is often called Occam's Razor, to choose among alternative explanations of their observations. The success of modern science lends credibility to the intuition. In terms of computational resources, shorter explanations have the advantage of using less memory than longer ones. The use of Occam's Razor for induction has been formalized and justified mathematically in several ways (Li and Vitanyi, 1992, and sources cited therein), but there is a substantial leap of faith in going from the mathematical ideal to real procedures for finding short explanations. The most convincing argument for Occam's Razor is still the powerful intuition behind it and the fact that it has served well in the past.

2.1 Evaluating Explanations

If a learning procedure is to choose the shortest explanation it must have some formal measure of explanation length. This section sketches some of the issues involved in developing such a measure, using concatenative morphology as an example.

Given a general theory such as the concatenative theory of word formation, the first step in developing a measure of explanation size is to choose a representation for explanations in terms of a finite set of symbols. The representation in Figure 1 is a first step, but it is not expressed entirely in terms of a finite symbolic alphabet. Some of the information in Figure 1 is represented by the distances between characters on the page and their alignment into rows and columns. One alternative would be to represent the explanations using one long string of keyboard characters, such as:



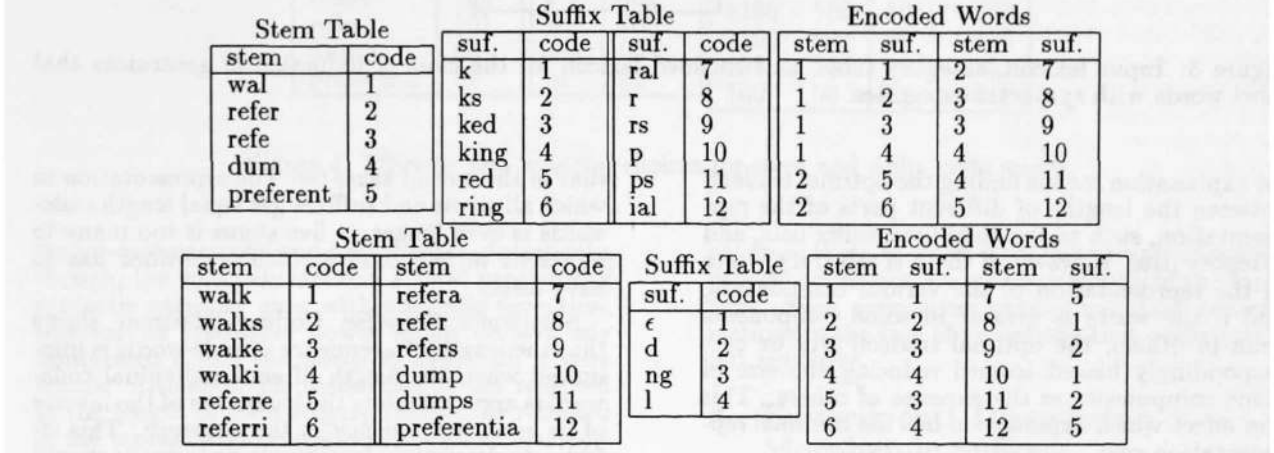
Input Words	
walk	referral
walks	refer
walked	refers
walking	dump
referred	dumps
referring	preferential

Stem Table	
stem	code
walk	1
referr	2
refer	3
dump	4
preferenti	5

Suffix Table	
suf.	code
ϵ	1
s	2
ed	3
ing	4
al	5

Encoded Words			
stem	suf.	stem	suf.
1	1	2	5
1	2	3	1
1	3	3	2
1	4	4	1
2	3	4	2
2	4	5	5

Figure 1: An input lexicon and a generative explanation for it.



Stem Table	
stem	code
wal	1
refer	2
refe	3
dum	4
preferent	5

Suffix Table			
suf.	code	suf.	code
k	1	ral	7
ks	2	r	8
ked	3	rs	9
king	4	p	10
red	5	ps	11
ring	6	ial	12

Encoded Words			
stem	suf.	stem	suf.
1	1	2	7
1	2	3	8
1	3	3	9
1	4	4	10
2	5	4	11
2	6	5	12

Stem Table			
stem	code	stem	code
walk	1	refera	7
walks	2	refer	8
walke	3	refers	9
walki	4	dump	10
referre	5	dumps	11
referri	6	preferentia	12

Suffix Table	
suf.	code
ϵ	1
d	2
ng	3
l	4

Encoded Words			
stem	suf.	stem	suf.
1	1	7	5
2	2	8	1
3	3	9	2
4	4	10	1
5	3	11	2
6	4	12	5

Figure 2: Generative explanations using suffixes that are too long (first line) and too short (second line).

```
walk-referr-refer-dump-preferenti:<e>-s-ed
-ing-al:1/1-1/2-1/3-...
```

If the conventions for converting between the table representation and the string are laid out clearly then the number of characters in the string can be taken as a measure of the *representation length* of the explanation. In fact, the number of characters could be calculated from the tables without actually constructing the string. This measure of representation length could then be used to decide which of two explanations is more plausible. Clearly, this measure would prefer the explanation in Figure 1 over those in Figure 2.

Although the string representation given above might well work for the problem of finding stems and suffixes, it does not put the best face on each explanation by representing it as briefly as possible. Further, the way in which the 41 characters are used leads to a certain arbitrariness in the representation length. For example, if the numbers in the last segment of the table get large enough to require several digits, it is wasteful to encode them in base 10, which uses only 10 of the 41 available symbols. A shorter encoding would use *a* for 10, *b* for 11, and so on. Conceivably, there might be two explanations, A and B, such that A is shorter when the numbers are encoded in base 10, and B is shorter when they are encoded in base 36.

This slight arbitrariness becomes more pronounced when more complex generative theories are considered. For example, the concatenation theory can be elaborated to account for the observed syntactic categories of words (assuming for the moment that the categories are known to the learner) as well as for their orthographic and phonetic forms. The theory would be that each suffix has an associated list of syntactic categories that are available to words ending in that suffix. Words are generated by picking a stem and a suffix as before, then assigning the word a syntactic category from the suffix's category list. An instantiation of this theory would include a stem list, a suffix list, and a category list for each suffix. The input sample would consist of words paired with their syntactic categories. (In this representation, ambiguous words occur once with each of their categories.) This schema is illustrated in Figure 3, assuming the same stem and suffix tables shown in Figure 1. The representation length of an explanation using an instantiation of this theory is a function of the lengths of the category lists as well as the stems and suffixes. For example, if the suffixes in some instantiation are too short, then they will not predict syntactic category accurately. For example, words ending in *-ify* are likely to be verbs, but words ending in *-y* may also be nouns (*day, bunny*), adjectives (*easy, funny*), adverbs (*quickly, slowly*), or prepositions (*by*). Thus, if *-y* is chosen as a morphemic suffix while *-ly* and *-ify* are not, the representation size of the category list will be longer. Often, finding the short-

A. Input Lexicon

walk	V	referring	V
walk	N	referring	A
walks	V	referral	N
walks	N	refer	V
walked	V	refers	V
walking	V	dog	N
walking	N	dogs	N
referred	V	preferential	A

B. Category Table

suf.	cat.	code
ϵ	V	0
	N	1
s	V	0
	N	1
ed	V	0
ing	V	0
	N	1
	A	2
al	A	0
	N	1

C. Encoded Lexicon

1	1	0	2	4	0
1	1	1	2	4	2
1	2	0	2	5	1
1	2	1	3	1	0
1	3	0	3	2	0
1	4	0	4	1	1
1	4	1	4	2	1
2	3	0	5	5	0

Figure 3: Input lexicon, category table, and encoded lexicon for the class of induction of generators that label words with syntactic categories.

est explanation means finding the optimal tradeoff between the lengths of different parts of the representation, such as the stem lists, suffix lists, and category lists. However, if there is arbitrary waste in the representation of the various components, and if the waste is greater in some components than in others, the optimal tradeoff will be correspondingly biased toward reducing the size of some components at the expense of others. This can affect which explanation has the minimal representation cost.

There is no way to ensure a complete absence of waste in a representation, but much of the obvious waste can be removed by standard changes of representation called *encodings*. The first step is to represent all portions of the explanation using the same alphabet. In view of the way current digital computers operate, an obvious choice is the two-symbol alphabet, whose symbols are usually written as 1 and 0. I will assume the binary alphabet, but in the final analysis the explanation chosen will not depend on the alphabet. The representation discussed above, using typewriter symbols, could be trivially converted to binary by choosing a unique binary sequence of six bits for each character, regardless of where it occurs in the representation. However, that would simply multiply the number of symbols in each representation by six, without eliminating any waste. At a minimum, integers should be represented in base two, rather than translating each decimal digit into a sequence of six bits. However, it is possible to go much further than that.

An exposition on coding theory would be out of place here (see, e.g. Hamming, 1986; Li and Vitanyi, 1992), but Figure 4 gives a general flavor for coding techniques. Rather than simply assigning sequential integers to the stems and suffixes, we can assign binary code words in such a way that the commonly used ones are shorter than the rarely used ones. An example is shown in Figure 4. In the top representation, more commonly used stems and suffix are assigned shorter code-words, whereas in the bottom representation the situation is reversed. The table of encoded words in the first representation takes only 60 bits, while that in the second takes 69. The representation in which all stems and suffixes get equal length code-words is even longer — five stems is too many to represent in two bits, so each code-word has to have three.

Shannon's Noiseless Coding Theorem shows that the length of a sequence of code words is minimized when the length of each individual code-word is approximately the logarithm of the inverse of its relative frequency in the sequence. This effectively determines how binary code words should be assigned to stems and suffixes. Let us use the term *symbol* to refer to the entities, such as the stems and suffixes, to which frequency-based code words are assigned. Then the representation length of the entire explanation is, to a first order approximation, the sum of the Shannon information of each of the symbol sets. (Further details are provided in Brent, Lundberg, and Murthy, 1993.) The Shannon Information is a simple function of the frequencies of the symbols — it can be computed without actually constructing binary code-words for each symbol. Thus, evaluating hypotheses is a computationally inexpensive procedure.

Figures 1 and 2 illustrate the strong intuitive connection between the number of characters in the stem and suffix lists of an instantiated theory and its predictive power. However, the importance of choosing stems and suffixes that minimize the size of the "Encoded Words" table is less obvious. To see why it is important, consider the refined theory that tries to explain the syntactic categories of words in terms of their morphemic analysis, as shown in Figure 3. The length each entry in the category table is a function of the number of different categories that words formed with the corresponding suffix belong to. Thus, if all words in the input ending in *-ly* are adverbs, the category table entry for *-ly* will be short, rewarding the choice of a suffix that predicts syntactic category. If, on the other hand, the input happens to contain the noun *bully*, and that word is erroneously analyzed as being formed with the morpheme *-ly*, then the length of the entry for *-ly* in the category table will nearly double. The entry would be the same length as if half the words end-

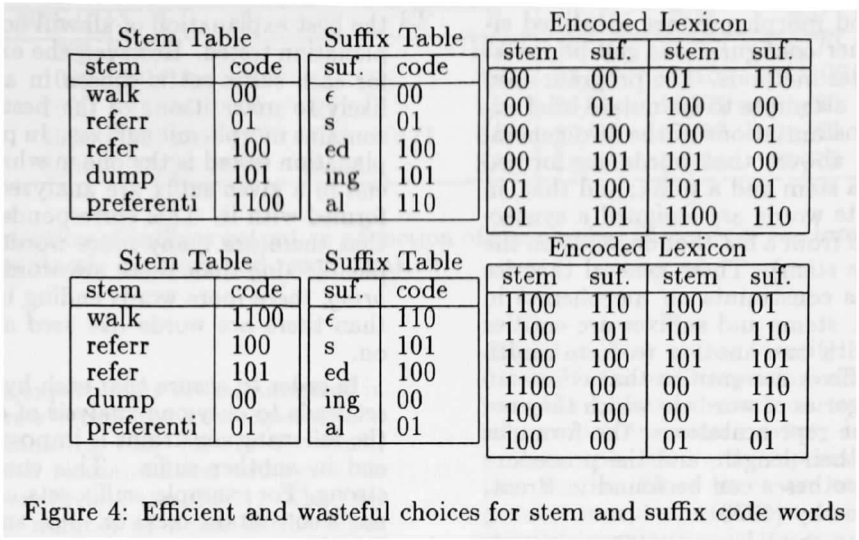


Figure 4: Efficient and wasteful choices for stem and suffix code words

ing in *-ly* were nouns and half adverbs. Intuitively, though, the suffix *-ly* remains a good predictor of syntactic category even with the *bully* exception, whereas its predictive power would be seriously diminished if half the words formed with it were nouns. Although this distinction is not reflected in the length of the category table, it is reflected in the representation length of the particular outcomes of the nondeterministic choices made by the generating procedure — i.e., in the "Encoded Words" table. The frequency-based encoding of symbols has the following property: among all sequences of n symbols drawn from a fixed symbol set, the more equal the frequencies of the symbols in the sequence the longer its encoded form; conversely, the more the sequence consists of a few common symbols plus some rare ones, the shorter its encoded form. For example, sequences of 99 copies of **Adverb** and one copy of **Noun** can be encoded much more briefly than sequences of 60 and 40, respectively. As a result, the cost of encoding the categories of words is greater if the words formed with *-ly* are equally likely to be adverbs or nouns than if they are almost all adverbs, with a few exceptions. This reflects the intuition that the misanalysis of *bully* should not be too damaging to the explanation that words formed with *-ly* are (generally) adverbs.

This example illustrates a key point about using representation length to evaluate hypothesis: it is robust in the face of error. A single misanalysis of *bully* as *bul-ly* has a small rather than a catastrophic effect on the evaluation of a hypothesis. Further, the evaluation is responsive to frequency. A few exceptions to a generalization do not necessarily cause the generalization to be rejected. Even if every suffix cooccurred occasionally with every syntactic category, the generalization that suffixes provide information about syntactic category might still reduce the cost of explanations that use the correct suffixes relative to those that use the wrong suffixes. As a result, learning procedures that exploit this evaluation are quite different from those in which a single example triggers a general conclusion about all examples of the same type. Such approaches cannot maintain generalizations in the face of exceptions that occur rarely but regularly.

3 Experiment: Discovering Morphemic Suffixes

This section describes an experiment aimed at answering the following questions about the informational structure of English:

1. To what extent can morphemes, which have consistent characteristics of form, syntax, and meaning, be discovered on the basis of form alone?
2. To what extent do non-morphemic regularities in the forms of words, such as the relatively high frequency of final *-sk* and *-ld* in English roots, interfere with the detection of morphemic regularities?
3. To what extent do adjustment rules, such as the deletion of final "e" before suffixes beginning in a vowel, obscure the forms of morphemes and interfere with the discovery of morphemic suffixes?
4. To what extent does the constraint between a composite word's final suffix and its major syntactic category help in identifying suffixes?

In this experiment, a computer program generates a series of hypotheses about the morphemic suffixes used to form words in an input sample and evaluates them by computing the representation length of the corresponding generative explanations. The input consists of words in standard orthography gleaned from journalistic text. The words are distinct, so this experiment tests a system of information flow in which the identification of whole words precedes the identification of morphemic suffixes. Of course, there are other interesting configurations, such as the one

in which words and morphemes are identified simultaneously. Other configurations can be investigated using similar methods. The program used in this experiment attempts to formulate brief explanations using instantiations of the two general theories described above: that words are formed by concatenating a stem and a suffix; and that, in addition, composite words are assigned a syntactic category chosen from a list that depends on the suffix (but not the stem). These general theories can be restated as constraints on morphemes in the following way: stems and suffixes are entities that recombine with one another to form multiple words; and suffixes are entities that constrain the syntactic categories of words in which they occur. Details of the representations, the formulas used to compute their length, and the procedure for generating hypotheses can be found in Brent, Lundberg, and Murthy (1993).

One point about child language acquisition is worth making at the outset: although natural languages have many non-concatenative morphological processes (Anderson, 1992), children may nonetheless have innate mechanisms of searching for concatenative morphemes. They must have some way of discovering non-concatenative morphology too, but the discovery process need not be homogeneous.

3.1 Generating Explanations

Methods for choosing among alternative explanations have been discussed, but the problem of generating explanations to choose among has not. Consider explanations for the forms of words, setting aside the question of syntactic categories. Each possible way of analyzing the words in some input into stems and suffixes constitutes a distinct explanation with its own representation size. For a word of length n, there are potentially n ways to analyze it. In order to limit the search to stem/suffix combinations, avoiding most prefix/stem combinations, only analyses in which the left half is at least as long as the right half are considered. Approximately $\frac{n}{2}$ analyses remain for a word of length n.

Changing the analysis of even a single word changes the explanation, so the number of possible explanations is the product, over all words in the input, of one-half their length. This number grows exponentially with the number of input words, so it is not possible to evaluate all of the explanations. However, the aim of this procedure is to discover morphemic suffixes, not to analyze individual words. Many explanations yield the same suffix set. For example, if the input contains *walking* and *string*, and if at least one of them is analyzed as having been formed with the suffix *-ing* in the shortest explanation, it is does not matter which one — *-ing* will be on the suffix list for that explanation in any case. The procedure tested here evaluates only one of the many possible explanations that correspond to each suffix set. Different explanations using the same suffix set may have different lengths, so it is possible that the best explanation of all will not be the best explanation tested. However, the explanation tested for each suffix set is chosen in a way that seems likely to make it one of the best if the suffix set contains morphemic suffixes. In particular, the explanation tested is the one in which all words that end in a given suffix are analyzed as having been formed with it. This corresponds to the intuition that there are many more words ending in morphemic *-ing* than there are words like *string* and *bring*, there more words ending in morphemic *-ed* than there are words like *seed* and *weed*, and so on.

In order to ensure that each hypothesized suffix set leads to only one analysis of each input word, the following constraint is imposed: no suffix may end in another suffix. This constraint is overly strong. For example, suffix sets containing *-s* cannot also contain *-ness* or *-ous*, and those containing *-ly* or *-ity* cannot also contain *-y*. There are various possibilities for relaxing this constraint, but they have not yet been tested. See Brent (1993) for additional heuristics governing the generation of hypotheses.

3.2 Methods

Input. The evaluation function and search techniques described above were tested on lexicons of various sizes. The lexicons were prepared from sample of the Wall Street Journal tagged for part-of-speech by the Penn Treebank project. All words except those containing capital letters or non-alphabetic symbols were sorted by frequency, and input lexicons of different sizes were prepared by taking the most common words from the top of the sorted list. Experiments were done using the theory of word form alone and the combined theory of word form and syntactic category. In the latter case, the Penn categories were mapped down to a set of five representing all common nouns, all verbs, all types of adjectives, all adverbs, and all other words.

Scoring. The results were scored in three major categories: the bound morphemes like *-ing* and *-ism*, the free morphemes in compounds like *-ball*, *-mark*, and *-man*, and the non-morphemic endings like "-ld" and "-sk". The reference for scoring bound morphemes was Marchand (1969) Independent words that appear as suffixes of another word and whose independent meaning contributes in any way to that of the whole were scored as free morphemes.

Some of the suffixes hypothesized by the system are clearly extensions of real morphemes. For example, one experiment yielded the suffix "-mental," as in *governmental*, which clearly contains the morpheme *-al*. Since these contain linguistically meaningful morphemes, they seem to have a different status than non-morphemic endings, such as the final "ld" of *mold, hold, held, build, sold, yield, bald, sold* etc. Extensions of bound and free morphemes were scored in their categories.

Words	B	EB	F	EF	E	Tot	%M	%P
500	6/7				10/3	16/10	38/70	38/70
1000	10/10				3/0	13/10	78/100	78/100
2000	15/16	2/1	2/2		1/1	20/20	95/95	85/90
4000	19/18	7/7	8/8	1/1	4/5	39/39	90/87	70/67
8000	27/28	9/9	19/18		3/6	58/61	95/90	79/75

Figure 5: Categorization of suffixes output as a function of the number of words in the input lexicon. Simple recombination / syntactic category and recombination.

3.3 Results

The results of the experiments are summarized in Figure 5. Each row represents two experiments: one using recombination only (above the slash) and one using both recombination and syntactic category prediction (below the slash). The first column shows the number of words in the input lexicon for each of the two experiments. The following five columns show the number of bound morphemes identified (B); extensions of bound morphemes (EB); free morphemes (F); extensions of free morphemes (EF); and errors (E), which includes all outputs that are not assigned to one of the other four categories. The final three columns show summary statistics: the total number of items hypothesized as morphemes (Tot); the percentage of those that were linguistically meaningful in some way, either as morphemes or extensions of morphemes (%M); and the percentage of the total that were perfect morphemes, neither extensions nor errors (%P).

For the procedure that attempts to explain the distribution of syntactic categories, the best explanation of the 1000 word lexicon used the following suffixes: *age al ed ing ion ity ly ment nce* and *s*. *nce* was counted as correct because it is the orthographic sequence common to a morphological process that yields either *ance*, as in *guidance*, or *ence*, as in *preference*. For the 2000 word lexicon the best explanation included all the suffixes from the 1000 word lexicon, plus the following new suffixes: *able ary ful ive ld ncy one out ship* and *sure*.

3.4 Discussion

The results of this experiment indicate that the formal combinatory properties of morphemic suffixes alone provide a tremendous amount of information about the stock of morphemes in English. The input lexicons contained thousands of non-morphemic endings and mere dozens of morphemic suffixes, but the output contained primarily morphemic suffixes in all cases but one. Thus, the effects of non-morphemic regularities are minimal. Adjustment rules had no noticeable effect. Thus, the "signal" from morphological recombination emerges quite clearly from the confusion of other signals, if the learning algorithm is tuned to it.

Consider first the effects of vocabulary size when syntactic categories are not used. When the input consists of only the 500 most common words in the corpus, the formal recombination property alone did not do very well at identifying morphemic suffixes. This is not surprising, since one would expect the 500 most common words in a journalistic corpus to contain a very high proportion of monomorphemic words, especially syntactic function words. Since there are so few instances of genuine morphemic suffixes to compete with, the spurious regularities in the corpus dominate. For input lexicons between 1000 and 8000 words the percentage of outputs that are genuine morphemes fluctuates around 80%, but shows no long-term trend.

Now consider the effect of adding in a syntactic property of morphemic suffixes — the fact that, in languages with right-headed words, the final suffix of a complex word predicts its major syntactic category. When explanations for the distribution of word-forms and their categories are represented as in Figure 3, category information appears to be a mixed blessing. For vocabularies of 2000 words or smaller it improves accuracy substantially, but when vocabulary size reaches 4000 to 8000 words, it shows a slight trend toward reducing accuracy. One possible explanation is that morphemic suffixes do not predict syntactic category very well in lower frequency words. A more plausible explanation, however, is that, since there are only five categories in this formulation, chance regularities are a significant factor by 8000 words. Syntactic information about words would probably be more useful if it were not limited to major categories. Since the effects of syntax are reflected in word-adjacency distributions, descriptive categories can be induced by clustering these distributions (Finch and Chater, 1992). Such categories might aid morpheme identification more than a priori major categories, since the induced categories would also reflect auxiliary structure, agreement, and related syntactic effects of morphology. In addition, there are more regularities to exploit even using the major categories. Notably, the category of the stem constrains the range of morphemic suffixes that can attach to it. Syntactic categories would probably be even more useful for an algorithm that exploited this regularity to help distinguish between morphemic suffixes and non-morphemic endings. In summary, this experiment suggests that knowing syntactic categories is marginally useful, but it does not suggest that they are any great bonanza. If they are not particularly useful for predicting syntactic category, then perhaps the flow of information should be primarily in the opposite direc-

tion, from morphemes to category assignment. On the other hand, morpheme discovery algorithms that make use of lexical syntactic information in more sophisticated ways may yet demonstrate that it is extremely valuable.

Presumably, it is most important for children to learn the most productive morphemes — those which speakers apply most freely to form words that have never heard or hear very rarely. Although many of the most productive English morphemes were discovered in these experiments, many low productivity morphemes were also discovered. Thus, it would be interesting to refine the algorithms used in the experiment so they focus more narrowly on the productive morphemes. One way to do this is to learn only from the lowest frequency words in the input, since the low frequency words are less likely to be memorized and more likely to be formed by productive morphological processes (Baayen and Lieber, 1991). At a minimum, this would have the benefit of eliminating the syntactic function words from consideration. It would also fit in nicely with an online processing model in which novel words draw the "attention" of the morpheme discovery procedures, whereas familiar words are quickly memorized and no longer subjected to analysis. However, experimentation with such a model awaits future work.

4 Conclusions

Minimum Generative Explanation algorithms are a promising tool for research in language acquisition, especially in domains like morphology and phonology. They appear to be the first algorithms that are robust enough to learn generative theories from naturally occurring input. If this research program is successful, it will ultimately yield a collection of techniques for discovering linguistic regularities in various domains. Taken together, these techniques will be capable of learning the regularities in typologically diverse languages. Beyond elucidating the questions of information flow, robustness, and search heuristics, they may also lead to models of how children modify their hypotheses incrementally, in response to individual inputs.

This research program emerges from a view of the language acquisition as a complex system governed by the interaction of loosely coupled mechanisms, each specialized for finding regularities of a particular type. On this view, the language acquisition device is analogous to a cell, which consists of a collection of complex molecules, each specialized to a particular task. The enzymes, for example, each carry out a particular chemical reaction, and these reactions are coupled in a dynamical system. Further, the behavior of each individual enzyme on a short timescale is stochastic, but its average behavior, stabilized by the behavior of other enzymes, is predictable, and the system as a whole is robust to a wide variety of inputs from the environment. This level of analysis contrasts with that of connectionist models, in which the individual elements are homogeneous, rather than specialized. It also contrasts with the level of analysis in parameter setting models and learnability theory, which look at the modes of behavior of the system as whole, rather than at the specialized elements from which it emerges.

References

Anderson, S. R. (1992). *A-Morphous Morphology.* Cambridge University Press, New York.

Baayen, H. and Lieber, R. (1991). Productivity and english derivation: A corpus-based study. *Linguistics*, 29:801–843.

Ellison, T. M. (1991). Discovering planar segregations. In *AAAI Spring Symposium on Machine Learning of Natural Language and Ontology.* AAAI.

Ellison, T. M. (1992). Learning vowel harmony. In *Background and Experiments in Machine Learning of Natural Language: Proceedings of the 1st* SHOE *Workshop.* Institute for Language Technology and Artificial Intelligence, Katholieke Universiteit, Brabant, Holland.

Finch, S. and Chater, N. (1992). Bootstrapping syntactic categories using statistical methods. In *Background and Experiments in Machine Learning of Natural Language: Proceedings of the 1st* SHOE *Workshop.* Institute for Language Technology and Artificial Intelligence, Katholieke Universiteit, Brabant, Holland.

Greenberg, J. H. (1966). *Some Universals of Grammar with Particular Reference to the Order of Meaningful Elements*, chapter 5, pages 73–113. MIT Press, Cambridge, MA.

Hamming, R. W. (1986). *Coding and Information Theory.* Prentice Hall, Englewood Cliffs NJ, second edition.

Li, M. and Vitányi, P. M. B. (1992). Inductive reasoning and kolmogorov complexity. *Journal of Computer and System Sciences*, 44:342–384.

Marchand, H. (1969). *The Categories and Types of Present-Day English Word-Formation.* C.H. Beck'sche Verlagsbuchhandlung, Munich.

Newport, E. and Singleton, J. (1993). Where learners surpass their models: The acquisition of asl from impoverished data. Manuscript, Dept. of Psychology, Rochester University.

Quinlan, J. R. and Rivest, R. L. (1989). Inferring decision trees using the minimum description length principle. *Information and Computation*, 80:227–248.

Rissanen, J. (1983). A universal prior for integers and estimation by minimum description length. *The Annals of Statistics*, 11:416–431.

Rissanen, J. (1986). Stochastistic complexity and sufficient statistics. Technical report.