From Grammar to Lexicon: Unsupervised Learning of Lexical Syntax

Michael R. Brent*
Johns Hopkins University

Imagine a language that is completely unfamiliar; the only means of studying it are an ordinary grammar book and a very large corpus of text. No dictionary is available. How can easily recognized, surface grammatical facts be used to extract from a corpus as much syntactic information as possible about individual words? This paper describes an approach based on two principles. First, rely on local morpho-syntactic cues to structure rather than trying to parse entire sentences. Second, treat these cues as probabilistic rather than absolute indicators of syntactic structure. Apply inferential statistics to the data collected using the cues, rather than drawing a categorical conclusion from a single occurrence of a cue. The effectiveness of this approach for inferring the syntactic frames of verbs is supported by experiments on an English corpus using a program called Lerner. Lerner starts out with no knowledge of content words—it bootstraps from determiners, auxiliaries, modals, prepositions, pronouns, complementizers, coordinating conjunctions, and punctuation.

1. Introduction

This paper presents a study in the automatic acquisition of lexical syntax from naturally occurring English text. It focuses on discovering the kinds of syntactic phrases that can be used to represent the semantic arguments of particular verbs. For example, *want* can take an infinitive argument and *hope* a tensed clause argument, but not vice versa:

(1) a. John wants Mary to be happy.
b. John hopes that Mary is happy.
c. *John wants that Mary is happy.
d. *John hopes Mary to be happy.

This study focuses on the ability of verbs to take arguments represented by infinitives, tensed clauses, and noun phrases serving as both direct and indirect objects. These lexical properties are similar to those that Chomsky (1965) termed *subcategorization frames*, but to avoid confusion the properties under study here will be referred to as *syntactic frames* or simply *frames*.

The general framework for the problems addressed in this paper can be thought of as follows. Imagine a language that is completely unfamiliar; the only means of studying it are an ordinary grammar book and a very large corpus of text (or transcribed speech). No dictionary is available. How can easily recognized, surface grammatical

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

facts be used to extract from a corpus as much syntactic information as possible about individual words?

The scenario outlined above is adopted in this paper as a framework for basic research in computational language acquisition. However, it is also an abstraction of the situation faced by engineers building natural language processing (NLP) systems for more familiar languages. The lexicon is a central component of NLP systems and it is widely agreed that current lexical resources are inadequate. Language engineers have access to some but not all of the grammar, and some but not all of the lexicon. The most easily formalized and most reliable grammatical facts tend to be those involving auxiliaries, modals, and determiners, the agreement and case properties of pronouns, and so on. These vary little from speaker to speaker, topic to topic, register to register. Unfortunately, this information is not sufficient to parse sentences completely, a fact that is underscored by the current state of the parsing art. If sentences cannot be parsed completely and reliably then the syntactic frames used in them cannot be determined reliably. How, then, can reliable, easily formalized grammatical information be used to extract syntactic facts about words from a corpus?

This paper suggests the following approach:

- Do not try to parse sentences completely. Instead, rely on local morpho-syntactic cues such as the following facts about English: (1) The word following a determiner is unlikely to be functioning as a verb; (2) The sequence *that the* typically indicates the beginning of a clause.
- Do not try to draw categorical conclusions about a word on the basis of one or a fixed number of examples. Instead, attempt to determine the distribution of exceptions to the expected correspondence between cues and syntactic frames. Use a statistical model to determine whether the cooccurrence of a verb with cues for a frame is too regular to be explained by randomly distributed exceptions.

The effectiveness of this approach for inferring the syntactic frames of verbs is supported by experiments using an implementation called Lerner. In the spirit of the problem stated above, Lerner starts out with no knowledge of content words—it bootstraps from determiners, auxiliaries, modals, prepositions, pronouns, complementizers, coordinating conjunctions, and punctuation. Lerner has two independent components corresponding to the two strategies listed above. The first component identifies sentences where a particular verb is likely to be exhibiting a particular syntactic frame. It does this using local cues, such as the *that the* cue. This component keeps track of the number of times each verb appears with cues for each syntactic frame as well as the total number of times each verb occurs. This process can be described as *collecting observations* and its output as an *observations table.* A segment of an actual observations table is shown in Table 4. The observations table serves as input to the statistical modeler, which ultimately decides whether the accumulated evidence that a particular verb manifests a particular syntactic frame in the input is reliable enough to warrant a conclusion.

To the best of my knowledge, this is the first attempt to design a system that autonomously learns syntactic frames from naturally occurring text. The goal of learning syntactic frames and the learning framework described above lead to three major differences between the approach reported here and most recent work in learning grammar from text. First, this approach leverages a little a priori grammatical knowledge using statistical inference. Most work on corpora of naturally occurring language either uses no a priori grammatical knowledge (Brill and Marcus 1992; Ellison 1991; Finch and Chater 1992; Pereira and Schabes 1992), or else it relies on a large and complex grammar (Hindle 1990, 1991). One exception is Magerman and Marcus (1991), in which a small grammar is used to aid learning.[1] A second difference is that the work reported here uses inferential rather than descriptive statistics. In other words, it uses statistical methods to infer facts about the language as it exists in the minds of those who produced the corpus. Many other projects have used statistics in a way that summarizes facts about the text but does not draw any explicit conclusions from them (Finch and Chater 1992; Hindle 1990). On the other hand, Hindle (1991) does use inferential statistics, and Brill (1992) recognizes the value of inference, although he does not use inferential statistics per se. Finally, many other projects in machine learning of natural language use input that is annotated in some way, either with part-of-speech tags (Brill 1992; Brill and Marcus 1992; Magerman and Marcus 1990) or with syntactic brackets (Pereira and Schabes 1992).

The remainder of the paper is organized as follows. Section 2 describes the morpho-syntactic cues Lerner uses to collect observations. Section 3 presents the main contribution of this paper—the statistical model and experiments supporting its effectiveness. Finally, Section 4 draws conclusions and lays out a research program in machine learning of natural language.

2. Collecting Observations

This section describes the local morpho-syntactic cues that Lerner uses to identify likely examples of particular syntactic frames. These cues must address two problems: finding verbs in the input and identifying phrases that represent arguments to the verb. The next two subsections present cues for these tasks. The cues presented here are not intended to be the last word on local cues to structure in English; they are merely intended to illustrate the feasibility of such cues and demonstrate how the statistical model accommodates their probabilistic correspondence to the true syntactic structure of sentences. Variants of these cues are presented in Brent (1991a, 1991b). The final subsection summarizes the procedure for collecting observations and discusses a sample of the observations table collected from the Brown corpus.

2.1 Finding Verbs

Lerner identifies verbs in two stages, each carried out on a separate pass through the corpus. First, strings that sometimes occur as verbs are identified. Second, occurrences of those strings in context are judged as likely or unlikely to be verbal occurrences. The second stage is necessary because of lexical ambiguity.

The first stage uses the fact that all English verbs can occur both with and without the suffix *-ing*. Words are taken as potential verbs if and only if they display this alternation in the corpus.[2] There are a few words that meet this criterion but do not occur as verbs, including *income/incoming* (*incame/incomed), *ear/earring*, *her/herring*, and *middle/middling*. However, the second stage of verb detection, combined with the statistical criteria, prevent these pairs from introducing errors.

[1] Brill and Marcus (1992) use a single grammatical rule in the test phase to supplement the rules their system learns, but no grammatical knowledge is used in the learning phase.

[2] Morphological analyzers typically use a root lexicon to resolve the ambiguities in morphological adjustment rules (Karttunen 1983). The system described here uses rules similar to those of Karttunen and Wittenburg (1983), but it resolves the ambiguities using only the contents of the corpus. This technique will be described in a subsequent paper.

Lerner assumes that a potential verb is functioning as a verb unless the context suggests otherwise. In particular, an occurrence of a potential verb is taken as a nonverbal occurrence only if it follows a determiner or a preposition other than *to*. For example, *was talking* would be taken as a verb, but *a talk* would not. This precaution reduces the likelihood that a singular count noun will be mistaken for a verb, since singular count nouns are frequently preceded by a determiner.

Finally, the only morphological forms that are used for learning syntactic frames are the stem form and the *-ing* form. There are several reasons for this. First, forms ending in *-s* are potentially ambiguous between third person singular present verbs and plural nouns. Since plural nouns are not necessarily preceded by determiners (*I like to take walks*), they could pose a significant ambiguity problem. Second, past participles do not generally take direct objects: *knows me* and *knew me* are OK, but not $*$ *is known me*. Further, the past tense and past participle forms of some verbs are identical, while those of others are distinct. As a result, using the *-ed* forms would have complicated the statistical model substantially. Since the availability of raw text is not generally a limiting factor, it makes sense to wait for the simpler cases.

2.2 Identifying Argument Phrases

When a putative occurrence of a verb is found, the next step is to identify the syntactic types of nearby phrases and determine whether or not they are likely to be arguments of the verb.

First, assume that a phrase P and a verb V have been identified in some sentence. Lerner's strategy for determining whether P is an argument to V has two components:

1. If P is a noun phrase (NP), take it as an argument only if there is evidence that it is not the subject of another clause.
2. Regardless of P's category, take it as an argument only if it occurs to the right of V and there are no potential attachment points for P between V and P.

For example, suppose that the sequence *that the* were identified as the left boundary of a clause in the sentence *I want to* **tell him that the** *idea won't fly*. Because pronouns like *him* almost never take relative clauses, and because pronouns are known at the outset, Lerner concludes that the clause beginning with *that the* is probably an argument of the verb *tell*.[3] It is always possible that it could be an argument of the previous verb *want*, but Lerner treats that as unlikely. On the other hand, if the sentence were *I want to* **tell the boss that the** *idea won't fly*, then Lerner cannot determine whether the clause beginning with *that the* is an argument to *tell* or is instead related to *boss*, as in *I want to fire the boss* **that the** *workers don't trust*.

Now consider specific cues for identifying argument phrases. The phrase types for which data are reported here are noun phrases, infinitive verb phrases (VPs), and tensed clauses. These phrase types yield three syntactic frames with a single argument and three with two arguments, as shown in Table 1. The cues used for identifying these frames are shown in Tables 2 and 3. Table 2 defines lexical categories that are referred to in Table 3. The category V in Table 3 starts out empty and is filled as verbs are detected on the first pass. "cap" stands for any capitalized word and "cap+" for any sequence of capitalized words. These cues are applied by matching them against the string of words immediately to the right of each verb. For example, a verb V is

[3] Thanks to Don Hindle for this observation (personal communication).

Table 1
The six syntactic frames studied in this paper.

SF Description	Good Example	Bad Example
NP only	greet them	*arrive them
tensed clause	hope he'll attend	*want he'll attend
infinitive	hope to attend	*greet to attend
NP & clause	tell him he's a fool	*yell him he's a fool
NP & infinitive	want him to attend	*hope him to attend
NP & NP	tell him the story	*shout him the story

Table 2
Lexical categories used in the definitions of the cues.

SUBJ:	I | he | she | we | they
OBJ:	me | him | us | them
SUBJ_OBJ:	you | it | yours | hers | ours | theirs
DET:	a | an | the | her | his | its | my | our | their | your | this | that | whose
+TNS:	has | have| had | am | is | are | was | were | do | does | did | can | could | may | might | must | will | would
CC:	when | before | after | as | while | if
PUNC:	. | ? | ! | , | ; | :

Table 3
Cues for syntactic frames. The category V is initially empty and is filled out during the first pass. "cap" stands for any capitalized word and "cap+" stands for any sequence of capitalized words.

Frame	Symbol	Cues
NP only	NP	(OBJ | SUBJ_OBJ | cap) (PUNC | CC)
Tensed Clause	cl	(that (DET | SUBJ | SUBJ_OBJ | cap+)) | SUBJ | (SUBJ_OBJ +TNS)
Infinitive VP	inf	to V
NP & clause	NPcl	(OBJ | SUBJ_OBJ | cap+) cl
NP & infinitive	NPinf	(OBJ | SUBJ_OBJ | cap+) inf
NP & NP (dat.)	NPNP	(OBJ | SUBJ_OBJ | cap+) NP

Table 4
A sample of the data collected from the untagged Brown Corpus using the cues of Table 3.

	V	NP	NPNP	NPcl	NPinf	cl	inf
recall	42	3				4	
recede	5						
receive	106	4					
reckon	10						
recognize	71	6				6	
recommend	32	2				1	
reconcile	5						
record	97	2					2
recount	5						
recover	14	4					1
recreate	2						
recruit	11						
recur	5						
redeem	3						
redirect	2						
rediscover	2						
reduce	85	2					
reek	2						
reel	2						
refer	43	1					1
refine	4						
reflect	41	1					
refresh	4						
refuel	3						
refuse	22	1					8

recorded as having occurred with a direct object and no other argument phrase if V is followed by a pronoun of ambiguous case and then a coordinating conjunction, as in *I'll* **see you when** *you return from Mexico.* The coordinating conjunction makes it unlikely that the pronoun is the subject of another clause, as in *I see you like champagne.* It also makes it unlikely that the verb has an additional NP argument, as in *I'll tell you my secret recipe.*

2.3 Summary and Sample Data

To summarize, the procedure for collecting observations from a corpus is as follows:

1. Go through the corpus once finding pairs of words such that one is the result of adding the suffix *-ing* to the other, applying appropriate morphological adjustment rules. List members of such pairs as verbs.
2. Go through the corpus again. At each word *w* that is on the list of verbs,
   (a) If *w* is not preceded by a preposition or a determiner, increment the number of times that *w* appears as a verb.
   (b) If any of the cues listed in Table 3 match the words immediately following *w*, increment the number of times that *w* appears to occur in the corresponding frame.
3. Combine the data for the stem form and the *-ing* form.

Table 4 shows an alphabetically contiguous portion of the observations table that results from applying this procedure to the Brown Corpus (untagged). Each row represents data collected from one pair of words, including both the *-ing* form and the stem form. The first column, titled V, represents the total number of times the word occurs in positions where it could be functioning as a verb. Each subsequent column represents a single frame. The number appearing in each row and column represents the number of times that the row's verb cooccurred with cues for the column's frame. Zeros are omitted. Thus *recall* and *recalling* occurred a combined total of 42 times, excluding those occurrences that followed determiners or prepositions. Three of those occurrences were followed by a cue for a single NP argument and four were followed by cues for a tensed clause argument.

Table 5
Judgments based on the observations in Table 4, made by the method of Section 3.

recall	NP, cl
recognize	NP, cl
recover	NP
refuse	inf

The cues are fairly rare, so verbs in Table 4 that occur fewer than 15 times tend not to occur with these cues at all. Further, these cues occur fairly often in structures other than those they are designed to detect. For example, *record, recover,* and *refer* all occurred with cues for an infinitive, although none of them in fact takes an infinitive argument. The sentences responsible for these erroneous observations are:

(2) (a) But I shall campaign on the Meyner **record** to meet the needs of the years ahead.
(b) Sposato needed a front, some labor stiff with a clean **record** to act as business agent of the Redhook local.
(c) Then last season the Birds tumbled as low as 11-18 on May 19 before **recovering** to make a race of it and total 86 victories.
(d) But I suspect that the old Roman was **referring** to change made under military occupation—the sort of change which Tacitus was talking about when....

In (2a,b) *record* occurs as a noun. In (2c) *recover* is a verb but the infinitive VP, *to make a race of it...*, does not appear to be an argument. In any case, it does not bear the same relation to the verb as the infinitive arguments of verbs like *try, want, hope, ask,* and *refuse*. In (2d) *refer* is a verb but *to change* is a PP rather than an infinitive.

The remainder of this paper describes and evaluates a method for making judgments about the ability of verbs to appear in particular syntactic frames on the basis of noisy data like that of Table 4. Given the data in Table 4, that method yields the judgments in Table 5.

3. Statistical Modeling

As noted above, the correspondence between syntactic structure and the cues that Lerner uses is not perfect. Mismatches between cue and structure are problematic because naturally occurring language provides no negative evidence. If a V verb is followed by a cue for some syntactic frame S, that provides evidence that V *does* occur in frame S, but there is no analogous source of evidence that V *does not* occur in frame S.

The occurrence of mismatches between cue and structure can be thought of as a random process where each occurrence of a verb V has some non-zero probability of being followed by a cue for a frame S, even if V cannot in fact occur in S. If this model is accurate, the more times V occurs, the more likely it is to occur at least once with a cue for S. The intransitive verb *arrive,* for example, will eventually occur with a cue for an NP argument, if enough text is considered. A learner that considers a single occurrence of verb followed by a cue to be conclusive evidence will eventually come to the false conclusion that *arrive* is transitive. In other words, the information provided by the cues will eventually be washed out by the noise. This problem is inherent in learning from naturally occurring language, since infallible parsing is not possible. The only way to prevent it is to consider the frequency with which each verb occurs with cues for each frame. In other words, to consider each occurrence of V *without* a cue for S as a small bit of evidence against V being able to occur in frame S. This section describes a statistical technique for weighing such evidence.

Given a syntactic frame S, the statistical model treats each verb V as analogous to a biased coin and each occurrence of V as analogous to a flip of that coin. An occurrence that is followed by a cue for S corresponds to one outcome of the coin flip, say heads; an occurrence without a cue for S corresponds to tails.[4] If the cues were perfect predictors of syntactic structure then a verb V that does not in fact occur in frame S would never appear with cues for S—the coin would never come up heads. Since the cues are not perfect, such verbs do occur with cues for S. The problem is to determine when a verb occurs with cues for S often enough that all those occurrences are unlikely to be errors.

In the following discussion, a verb that in fact occurs in frame S in the input is described as a $+S$ verb; one that does not is described as a $-S$ verb. The statistical model is based on the following approximation: for fixed S, all $-S$ verbs have equal probability of being followed by a cue for S. Let π_{-s} stand for that probability. π_{-s} may vary from frame to frame, but not from verb to verb. Thus, errors might be more common for tensed clauses than for NPs, but the working hypothesis is that all intransitives, such as *saunter* and *arrive*, are about equally likely to be followed by a cue for an NP argument. If the error probability π_{-s} were known, then we could use the standard hypothesis testing method for binomial frequency data. For example, suppose $\pi_{-s} = .05$—on average, one in twenty occurrences of a $-S$ verb is followed by a cue for S. If some verb V occurs 200 times in the corpus, and 20 of those occurrences are followed by cues for S, that ought to suggest that V is unlikely to have probability .05 of being followed by a cue for S, and hence V is unlikely to be $-S$. Specifically, the chance of flipping 20 or more heads out of 200 tosses of a coin with a five percent chance of coming up heads each time is less than three in 1000. On the other hand, it is not all that unusual to flip 2 or more heads out of 20 on such a coin—it happens about one time in four. If a verb occurs 20 times in the corpus and 2 of those occurrences are followed by cues for S, it is quite possible that V is $-S$ and that the 2 occurrences with cues for S are explained by the five percent error rate on $-S$ verbs.

The next section reviews the hypothesis-testing method and gives the formulas for computing the probabilities of various outcomes of coin tosses, given the coin's bias. It also provides empirical evidence that, for some values of π_{-s}, hypothesis-testing does a good job of distinguishing $+S$ verbs from $-S$ verbs that occur with cues for S because of mismatches between cue and structure. The following section proposes a method for estimating π_{-s} and provides empirical evidence that its estimates are nearly optimal.

3.1 Hypothesis Testing

The statistical component of Lerner is designed to prevent the information provided by the cues from being washed out by the noise. The basic approach is hypothesis testing on binomial frequency data (Kalbfleisch 1985). Specifically, a verb V is shown to

[4] Given a verb V, the outcomes of the coins for different S's are treated as approximately independent, even though they cannot be perfectly independent. Their dependence could be modeled using a multinomial rather than a binomial model, but the experimental data suggest that this is unnecessary.

be $+S$ by assuming that it is $-S$ and then showing that if this were true, the observed pattern of cooccurrence of V with cues for S would be extremely unlikely.

3.1.1 Binomial Frequency Data. In order to use the hypothesis testing method we need to estimate the probability π_{-s} that an occurrence of a verb V will be followed by a cue for S *if V is* $-S$. In this section it is assumed that π_{-s} is known. The next section suggests a means of estimating π_{-s}. In both sections it is also assumed that for each $+S$ verb, V, the probability that V will be followed by a cue for S is greater than π_{-s}. Other than that, no assumptions are made about the probability that a $+S$ verb will be followed by a cue for S. For example, two verbs with transitive senses, such as *cut* and *walk*, may have quite different frequencies of cooccurrence with cues for NP. It does not matter what these frequencies are as long as they are greater than π_{-NP}.

If a coin has probability p of flipping heads, and if it is flipped n times, the probability of its coming up heads exactly m times is given by the binomial distribution:

$$P(m,n,p) = \frac{n!}{m!(n-m)!}p^m(1-p)^{n-m} \tag{1}$$

The probability of coming up heads *m or more* times is given by the obvious sum:

$$P(m+,n,p) = \sum_{i=m}^{n} P(i,n,p) \tag{2}$$

Analogously, $P(m+,n,\pi_{-s})$ gives the probability that m or more occurrences of a $-S$ verb V will be followed by a cue for S out of n occurrences total.

If m out of n occurrences of V are followed by cues for S, and if $P(m+,n,\pi_{-s})$ is quite small, then it is unlikely that V is $-S$. Traditionally, a threshold less than or equal to .05 is set such that a hypothesis is rejected if, assuming the hypothesis were true, the probability of outcomes as extreme as the observed outcome would be below the threshold. The confidence attached to this conclusion increases as the threshold decreases.

3.1.2 Experiment. The experiment presented in this section is aimed at determining how well the method presented above can distinguish $+S$ verbs from $-S$ verbs. Let p_{-s} be an estimate of π_{-s}. It is conceivable that $P(m+,n,p_{-s})$ might not be a good predictor of whether or not a verb is $+S$, regardless of the estimate p_{-s}. For example, if the correspondence between the cues and the structures they are designed to detect were quite weak, then many $-S$ verbs might have lower $P(m+,n,p_{-s})$ than many $+S$ verbs. This experiment measures the accuracy of binomial hypothesis testing on the data collected by Lerner's cues as a function of p_{-s}. In addition to showing that $P(m+,n,p_{-s})$ is good for distinguishing $+S$ and $-S$ verbs, these data provide a baseline against which to compare methods for estimating the error rate π_{-s}.

Method The cues described in Section 2 were applied to the Brown Corpus (untagged version). Equation 2 was applied to the resulting data with a cutoff of $P(m+,n,p_{-s}) < .02$ and p_{-s} varying between 2^{-5} (1 error in every 32 occurrences) and 2^{-13} (1 error in every 8192 occurrences). The resulting judgments were compared to the blind judgments of a single judge. One hundred ninety-three distinct verbs were chosen at random from the tagged version of the Brown Corpus for comparison. Common verbs are more likely to be included in the test sample than rare verbs, but

Table 6
Comparison of automatic classification to hand judgments for tensed-clause complement as a function of estimated error rate p (Brown Corpus). PRE = (TP / TP + FP); REC = (TP / TP + FN).

$-\log_2 p_{-cl}$	p_{-cl}	TP	FP	TN	FN	MC	%MC	PRE	REC
5	.0312	13	0	30	20	20	32	1.00	.39
6	.0156	19	0	30	14	14	22	1.00	.58
7	.0078	22	1	29	11	12	19	.96	.67
8	.0039	25	1	29	8	9	14	.96	.76
9	.0020	27	3	27	6	9	14	.90	.82
10	.0010	29	5	25	4	9	14	.85	.88
11	.0005	31	8	22	2	10	16	.79	.94
12	.0002	31	13	17	2	15	24	.70	.94
13	.0001	33	19	11	0	19	30	.63	1.00

no verb is included more than once. Each verb was scored for a given frame only if it cooccurs with a cue for that frame at least once. Thus, although 193 verbs were randomly selected from the corpus for scoring, only the 63 that cooccur with a cue for tensed clause at least once were scored for the tensed-clause frame. This procedure makes it possible to evaluate the hypothesis-testing method on data collected by the cues, rather than evaluating the cues per se. It also makes the judgment task much easier—it is not necessary to determine whether a verb can appear in a frame in principle, only whether it does so in particular sentences. There were, however, five cases where the judgments were unclear. These five were not scored. See Appendix C for details.

Results The results of these comparisons are summarized in Table 6 (tensed clause) and Table 7 (infinitive). Each row shows the performance of the hypothesis-testing procedure for a different estimate p_{-s} of the error-rate π_{-s}. The first column shows the negative logarithm of p_{-s}, which is varied from 5 (1 error in 32 occurrences) to 13 (1 error in 8192 occurrences). The second column shows p_{-s} in decimal notation. The next four columns show the number of *true positives* (TP)—verbs judged $+S$ both by machine and by hand; *false positives* (FP)—verbs judged $+S$ by machine, $-S$ by hand; true negatives (TN)—verbs judged $-S$ both by machine and by hand; and false negatives (FN)—verbs judged $-S$ by machine, $+S$ by hand. The numbers represent distinct verbs, not occurrences. The seventh column shows the number of verbs that were *misclassified* (MC)—the sum of false positives and false negatives. The eighth column shows the *percentage* of verbs that were *misclassified* (%MC). The next-to-last column shows the *precision* (PRE)—the true positives divided by all verbs that Lerner judged to be $+S$. The final column shows the *recall* (REC)—the true positives divided by all verbs that were judged $+S$ by hand.

Discussion For verbs taking just a tensed clause argument, Table 6 shows that, given the right estimate p_{-s} of π_{-s}, it is possible to classify these 63 verbs with only 1 false positive and 8 false negatives. If the error rate were ignored or approximated as zero then the false positives would go up to 19. On the other hand, if the error rate were taken to be as high as 1 in 2^5 then the false negatives would go up to 20. In this case, the sum of both error types is minimized with $2^{-8} \leq p_{-cl} \leq 2^{-10}$. Table 7 shows similar results for verbs taking just an infinitive argument, where misclassifications are minimized with $p_{-inf} = 2^{-7}$.

Table 7
Comparison of automatic classification to hand judgments for infinitive complement, as a function of estimated error rate p (Brown Corpus).

$-\log_2 p_{-inf}$	p_{-inf}	TP	FP	TN	FN	MC	%MC	PRE	REC
5	.0312	14	0	33	13	13	22	1.00	.52
6	.0156	16	0	33	11	11	18	1.00	.59
7	.0078	19	1	32	8	9	15	.95	.70
8	.0039	22	6	27	5	11	18	.79	.81
9	.0020	22	8	25	5	13	22	.73	.81
10	.0010	24	12	21	3	15	25	.67	.89
11	.0005	24	14	19	3	17	28	.63	.89
12	.0002	26	19	14	1	20	33	.58	.96
13	.0001	27	26	7	0	26	43	.51	1.00

3.2 Estimating the Error Rate

As before, assume that an occurrence of a $-S$ verb is followed by a cue for S with probability π_{-s}. Also as before, assume that for each $+S$ verb V, the probability that an occurrence of V is followed by a cue for S is greater than π_{-s}.

It is useful to think of the verbs in the corpus as analogous to a large bag of coins with various biases, or probabilities of coming up heads. The only assumption about the distribution of biases is that there is some definite but unknown minimum bias π_{-s}.[5] Determining whether or not a verb appears in frame S is analogous to determining, for some randomly selected coin, whether its bias is greater than π_{-s}. The only available evidence comes from selecting a number of coins at random and flipping them. The previous section showed how this can be done given an estimate of π_{-s}.

Suppose a series of coins is drawn at random from the bag. Each coin is flipped N times. It is then assigned to a histogram bin representing the number of times it came up heads. At the end of this sampling procedure bin i contains the number of coins that came up heads exactly i times out of N. Such a histogram is shown in Figure 1, where $N = 40$. If N is large enough and enough coins are flipped N times, one would expect the following:

1. The coins whose probability of turning up heads is π_{-s} (the minimum) should cluster at the low-heads end of the histogram. That is, there should be some $0 \leq j_0 \leq N$ such that most of the coins that turn up j_0 heads or fewer have probability π_{-s}, and, conversely, most coins with probability π_{-s} turn up j_0 heads or fewer.

2. Suppose j_0 were known. Then the portion of the histogram below j_0 should have a roughly binomial shape. In Figure 1, for example, the first eight bins have roughly the shape one would expect if j_0 were 8. In contrast, the first 16 bins do not have the shape one would expect if j_0

[5] If the number of coins is taken to be infinite, then the biases must be not only greater than π_{-s} but bounded above π_{-s}.

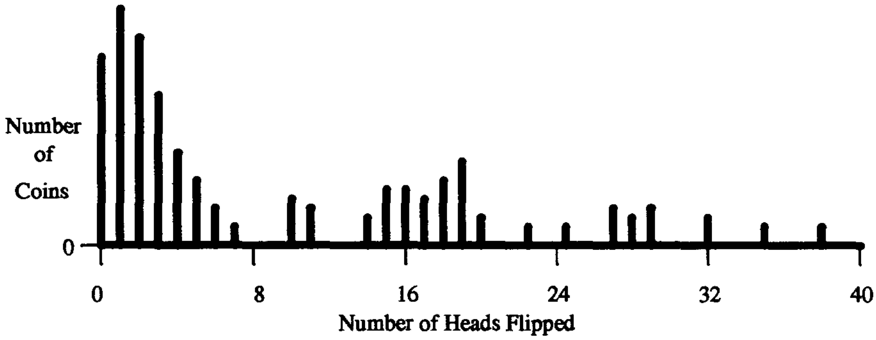

Figure 1
A histogram illustrating a binomially shaped distribution in the first eight bins.

were 16—their height drops to zero for two stretches before rising significantly above zero again. Specifically, the height of the i^{th} histogram bin should be roughly proportional to $P(i, N, p_{-s})$, with N the fixed sample size and p_{-s} an estimate of π_{-s}.

3. Suppose again that j_0 were known. Then the average rate at which the coins in bins j_0 or lower flip heads is a good estimate of π_{-s}.

The estimation procedure tries out each bin as a possible estimate of j_0. Each estimate of j_0 leads to an estimate of π_{-s} and hence to an expected shape for the first j_0 histogram bins. Each estimate j of j_0 is evaluated by comparing the predicted distribution in the first j bins to the observed distribution—the better the fit, the better the estimate.

Moving from coins to verbs, the procedure works as follows. For some fixed N, consider the first N occurrences of each verb that occurs at least N times in the input. (A uniform sample size N is needed only for estimating π_{-s}. Given an estimate of π_{-s}, verbs with any number of occurrences can be classified.) Let S be some syntactic frame and let $H[i]$ be the number of distinct verbs that were followed by cues for S exactly i times out of N—i.e., the height of the ith histogram bin. Assume that there is some $1 \leq j_0 \leq N$ such that most $-S$ verbs are followed by cues for S j_0 times or fewer, and conversely most verbs that are followed by cues for S j_0 times or fewer are $-S$ verbs. For each possible estimate j of j_0 there is a corresponding estimate of π_{-s}; namely, the average rate at which verbs in the first j bins are followed by cues for S. Choosing the most plausible estimate of π_{-s} thus comes down to choosing the most plausible estimate of j_0, the boundary between the $-S$ verbs and the rest of the histogram. To evaluate the plausibility of each possible estimate j of j_0, measure the fit between the predicted distribution of $-S$ verbs, assuming j is the boundary of the $-S$ cluster, and the observed distribution of the $-S$ verbs, also assuming j is the boundary of the $-S$ cluster. Given j, let p_{-s} stand for the average rate at which verbs in bins j or lower are followed by cues for S. The predicted distribution for $-S$ verbs is proportional to $P(i, N, p_{-s})$ for $0 \leq i \leq N$. The observed distribution of $-S$ verbs, assuming j is the boundary of the $-S$ cluster, is $H[i]$ for $0 \leq i \leq j$ and 0 for $j < i \leq N$. Measure the fit between the predicted and observed distributions by normalizing both to have unit area and taking the sum over $0 \leq i \leq N$ of the squares of the differences between the two distributions at each bin i.

Table 8
Comparison of automatic classification using the Brown Corpus to hand judgments. The estimate p_{-S} is made with $N = 100$. The probability threshold is .02.

S	j	p_{-s}	TP	FP	TN	FN	MC	%MC	PRE	REC
cl	2	0.0037	25	1	28	8	9	15	.96	.76
inf	2	0.0048	22	1	32	5	6	10	.96	.81
NPcl	1	0.0002	3	2	2	0	2	29	.60	1.00
NPinf	1	0.0005	5	0	3	2	2	20	1.00	.71
NPNP	3	0.0004	3	0	3	3	3	33	1.00	.50
NP	4	0.0132	52	1	5	59	60	51	.98	.47
total			110	5	73	74	79	30	.96	.60

In pseudo-code, the procedure is as follows:

```
ESTIMATE-P(H[], N)
area := $H[0]$, min-sum-of-squares := $\infty$, best-estimate := 1;
Try each value of j from 1 to N as an estimate of $j_0$
for $j$ from 1 to $N$
    $p_{-s}$ := 0
    area := area + $H[j]$
    for $i$ from 0 to j
        Normalize the $-S$ bins to area 1.
        $H'[i]$ := $\frac{H[i]}{\text{area}}$
        Estimate $\pi_{-s}$ by the average cooccurrence rate for
        the first j bins—those presumed to hold $-S$ verbs
        $p_{-s}$ := $p_{-s} + (\frac{i}{N} * H'[i])$
    Check the fit, assuming j is the $\pm S$ boundary
    sum-of-squares := 0
    for $i$ from 0 to N
        Compute the predicted distribution for bin i
        $P$ := $\frac{N!}{i!(N-i)!} p_{-s}^{i} (1 - p_{-s})^{N-i}$
        Verbs in the first bins j and below are presumed $-S$
        if $i \leq j$
        then normalized-observed := $H'[i]$
        else normalized-observed := 0
        sum-of-squares := sum-of-squares $+$ (normalized-observed $-$ $P)^2$
    Choose the $p_{-s}$ yielding the best fit
    if sum-of-squares $<$ min-sum-of-squares
    then min-sum-of-squares := sum-of-squares
         best-estimate := $p_{-s}$
return best-estimate
```

3.2.1 Experiment. This section evaluates the proposed estimation technique empirically in terms of the errors it yields when the cues of Section 2 are applied to the Brown Corpus. The sample selection and scoring procedures are the same as in the previous section. When π_{-s} is estimated using sample size $N = 100$, Table 8 shows the results for each of the six frames. Varying N between 50 and 150 results in no significant change in the estimated error rates.

One way to judge the value of the estimation and hypothesis-testing methods is to examine the false positives. Three of the five false positives result from errors in verb detection that are not distributed uniformly across verbs. In particular, *shock, board,* and *near* are used more often as nonverbs than as verbs. This creates many opportunities for nonverbal occurrences of these words to be mistaken for verbal occurrences. Other verbs, like *know,* are unambiguous and thus are not subject to this type of error. As a result, these errors violate the model's assumption that errors are distributed uniformly across verbs and highlight the limitations of the model. The remaining false positives were *touch* and *belong,* both mistaken as taking an NP followed by a tensed clause. The *touch* error was caused by the capitalization of the first word of a line of poetry:

> I knew not what did to a friend *belong*
> *Till I* stood up, true friend, by thy true side;

Till was mistaken for a proper name. The *belong* error was caused by mistaking a matrix clause for an argument in:

> With the blue flesh of night *touching him he* stood under a gentle hill caressing the flageolet with his lips, making it whisper.

It seems likely that such input would be much rarer in more mundane sources of text, such as newspapers of record, than in the diverse Brown Corpus.

The results for infinitives and clauses can also be judged by comparison to the optimal classifications rates from Tables 6 and 7. In both cases the classification appears to be right in the optimal range. In fact, the estimated error rate for infinitives produces a better classification than any of those shown in Table 7. (It falls at a value between those shown.) The classification of clauses and infinitives remains in the optimal range when the probability threshold is varied from .01 to .05.

Overall the tradeoff between improved precision and reduced recall seems quite good, as compared to doing no noise reduction ($p_{-s} = 0$). The only possible exception is the NP frame, where noise reduction causes 59 false negatives in exchange for preventing only 5 false positives. This is partly explained by the different prior probabilities of the different frames. Most verbs can take a direct object argument, whereas most verbs cannot take a direct object argument followed by a tensed clause argument. There is no way to know this in advance. There may be other factors as well. If the error rate for the NP cues is substantially lower than 1 out of 100, then it cannot be estimated accurately with sample size $N = 100$. On the other hand, if the sample size N is increased substantially there may not be enough verbs that occur N times or more in the corpus. So a larger corpus might improve the recall rate for *NP.*

4. General Discussion

This paper explores the possibility of using simple grammatical regularities to learn lexical syntax. The data presented in Tables 6, 7, and 8 provide evidence that it is possible to learn significant aspects of English lexical syntax in this way. Specifically, these data suggest that neither a large parser nor a large lexicon is needed to recover enough syntactic structure for learning lexical syntax. Rather, it seems that significant lexical syntactic information can be recovered using a few approximate cues along with statistical inference based on a simple model of the cues' error distributions.

4.1 Other Syntactic Frames

The lexical entry of a verb can specify other syntactic frames in addition to the six studied here. In particular, many verbs take prepositional phrases (PPs) headed by a particular preposition or class of prepositions. For example, *put* requires a location as a second argument, and locations are often represented by PPs headed by locative prepositions.

Extending Lerner to detect PPs is trivial. Since the set of prepositions in the language is essentially fixed, all prepositions can be included in the initial lexicon. Detecting a PP requires nothing more than detecting a preposition.[6] The statistical model can, of course, be applied without modification.

The problem, however, is determining which PPs are arguments and which are adjuncts. There are clearly cases where a prepositional phrase can occur in a clause not by virtue of the lexical entry of the verb but rather by virtue of nonlexical facts of English syntax. For instance, almost any verb can occur with a temporal PP headed by *on,* as in *John arrived on Monday.* Such PPs are called *adjuncts.* On the other hand, the sense of *on* in *John sprayed water on the ceiling* is quite different. This sense, it can be argued, is available only because the lexical entry of *spray* specifies a location argument that can be realized as a PP. If anything significant is to be learned about individual words, the nonspecific cooccurrences of verbs with PPs (adjuncts) must be separated from the specific ones (arguments). It is not clear how a machine learning system could do this, although frequency might provide some clue. Worse, however, there are many cases in which even trained linguists lack clear intuitions. Despite a number of attempts to formulate necessary and sufficient conditions for the argument/adjunct distinction (e.g., Jackendoff 1977), there are many cases for which the various criteria do not agree or the judgments are unclear (Adams and Macfarland 1991). Thus, the Penn Treebank does not make the argument/adjunct distinction because their judges do not agree often enough. Until a useful definition that trained humans can agree on is developed, it would seem fruitless to attempt machine learning experiments in this domain.

4.2 Limitations of the Statistical Model

Although the results of this study are generally encouraging, they also point to some limitations of the statistical model presented here. First, it does not take into account variation in the percentage of verbs that can appear in each frame. For example, most verbs can take an NP argument, while very few can take an NP followed by a tensed clause. This results in too few verbs being classified as +NP and too many being classified as +NPcl, as shown in Table 8. Second, it does not take into account the fact that for some words with verbal senses most of their occurrences are verbal, whereas for others most of their occurrences are nonverbal. For example, *operate* occurs exclusively as a verb while *board* occurs much more often as a noun than as a verb. Since the cues are based on the assumption that the word in question is a verb, *board* presents many more opportunities for error than *operate.* This violates the assumption that the probability of error for a given frame is approximately uniform across verbs.

[6] The preposition/particle distinction is set aside here in order to focus on the more problematic argument/adjunct distinction.

Table 9
Distribution of occurrences among morphological forms in the Brown Corpus. The ambiguous words *board* and *project* show a pattern of distribution distinct from that of the unambiguous verbs *operate* and *follow*.

project	52	board	111	operate	48	follow	76
projects	54	boards	31	operates	15	follows	72
projected	10	boarded	3	operated	26	followed	150
projecting	5	boarding	1	operating	55	following	97

These limitations do not constitute a major impediment to applications of the current results. For example, an applied system can be provided with the rough estimates that 80–95 percent of verbs take a direct object, while 1–2 percent take a direct object followed by a tensed clause. Such estimates can be expected to reduce misclassification significantly. Further, an existing dictionary could be used to "train" a statistical model on familiar verbs. A trained system would probably be more accurate in classifying new verbs. Finally, the lexical ambiguity problem could probably be reduced substantially in the applied context by using a statistical tagging program (Brill 1992; Church 1988).

For addressing basic questions in machine learning of natural language the solutions outlined above are not attractive. All of those solutions provide the learner with additional specific knowledge of English, whereas the goal for the machine learning effort should be to replace specific knowledge with general knowledge about the types of regularities to be found in natural language.

There is one approach to the lexical ambiguity problem that does not require giving the learner additional specific knowledge. The problem is as follows: words that occur frequently as, say, nouns are likely to have a different error rate from unambiguous verbs. If it were known which words occur primarily as verbs and which occur primarily as nouns then separate error rate estimates could be made for each. This would reduce the rate of false positive errors even without any further information about which *particular* occurrences are nominal and which are verbal. One way to distinguish primarily nominal words from primarily verbal words is by the relative frequencies of their various inflected forms. For example, Table 9 shows the contrast in the distribution of inflected forms between *project* and *board* on the one hand and *operate* and *follow* on the other. *Project* and *board* are two words whose frequent occurrence as nouns has caused Lerner to make false positive errors. In both cases, the stem and *-s* forms are much more common than the *-ed* and *-ing* forms. Compare this to the distribution for the unambiguous verbs *operate* and *follow*. In these cases the diversity of frequencies is much lower and does not display the characteristic pattern of a word that occurs primarily as a noun— *-ing* and *-ed* forms that are much rarer than the *-s* and stem forms. Similar characteristic patterns exist for words that occur primarily as adjectives. Recognizing such ambiguity patterns automatically would allow a separate error rate to be estimated for the highly ambiguous words.

4.3 Future Work

From the perspective of computational language acquisition, a natural direction in which to extend this work is to develop algorithms for learning some of the specific knowledge that was programmed into the system described above. Consider the morphological adjustment rules according to which, for example, the final "e" of *bite* is deleted when the suffix *-ing* is added, yielding *biting* rather than *"biteing." Lerner needs to know such rules in order to determine whether or not a given word occurs both with and without the suffix *-ing*. Experiments are under way on an unsupervised procedure that learns such rules from English text, given only the list of English verbal suffixes. This work is being extended further in the direction of discovering the morphemic suffixes themselves and discovering the ways in which these suffixes alternate in paradigms. The short-term goal is to develop algorithms that can learn the rules of inflection in English starting from only a corpus and a general notion of the nature of morphological regularities.

Ultimately, this line of inquiry may lead to algorithms that can learn much of the grammar of a language starting with only a corpus and a general theory of the kinds of formal regularities to be found in natural languages. Some elements of syntax may not be learnable in this way (Lightfoot 1991), but the lexicon, morphology, and phonology together make up a substantial portion of the grammar of a language. If it does not prove possible to learn these aspects of grammar starting from a general ontology of linguistic regularities and using distributional analysis then that, too, is an interesting result. It would suggest that the task requires a more substantive initial theory of possible grammars, or some semantic information about input sentences, or both. In any case this line of inquiry promises to shed light on the nature of language, learning, and language learning.

References

Adams, L., and Macfarland, T. (1991). "Testing for adjuncts." In *Proceedings, Second Annual Meeting of the Formal Linguistics Society of Midamerica.* Formal Linguistics Society of Midamerica.

Brent, M. R. (1991a). *Automatic acquisition of subcategorization frames from unrestricted English.* Doctoral dissertation, Massachusetts Institute of Technology, Cambridge MA.

Brent, M. R. (1991b). "Automatic acquisition of subcategorization frames from untagged text." In *Proceedings, 29th Annual Meeting of the ACL.* 209–214.

Brill, E. (1992). "A simple rule-based part of speech tagger." In *Darpa Workshop on Speech and Natural Language.* Harriman.

Brill, E., and Marcus, M. (1992). "Automatically acquiring phrase structure using distributional analysis." In *Darpa Workshop on Speech and Natural Language.* Harriman.

Chomsky, N. (1965). *Aspects of the Theory of Syntax.* MIT Press.

Church, K. (1988). "A stochastic parts program and noun phrase parser for unrestricted text." In *Proceedings, Second ACL Conference on Applied NLP.*

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

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.* Katholieke Universiteit, Brabant, Holland.

Hindle, D. (1990). "Noun classification from predicate argument structures." In *Proceedings, 28th Annual Meeting of the ACL.* 268–275.

Hindle, D. (1991). "Structural ambiguity and lexical relations." In *Proceedings, 29th Annual Meeting of the ACL.* 229–236.

Jackendoff, R. (1977). *X-bar Syntax: A Study of Phrase Structure.* Volume 2 of *Linguistic Inquiry Monograph.* MIT Press.

Kalbfleisch, J. G. (1985). *Probability and Statistical Inference,* Volume 2, Second Edition. Springer-Verlag.

Karttunen, L. (1983). "KIMMO: A general morphological processor." *Texas Linguistics Forum,* 22(22), 165–186.

Karttunen, L., and Wittenburg, K. (1983). "A two-level morphological analysis of English." *Texas Linguistics Forum,* 22(22), 217–223.

Lightfoot, D. W. (1991). *How to Set Parameters.* MIT Press.

Magerman, D., and Marcus, M. (1990). "Parsing a natural language using mutual information statistics." In *Proceedings, Eighth National Conference on Artificial Intelligence.*

Magerman, D., and Marcus, M. (1991). "Distituent parsing and grammar induction." In *AAAI Spring Symposium on Machine Learning of Natural Language and Ontology.*

Pereira, F., and Schabes, Y. (1992). "Inside-outside reestimation from partially bracketed corpora." In *Proceedings, 30th Annual Meeting of the Association for Computational Linguistics.*

Appendix A: Test Words

The experiments described above used the following 193 verbs, selected at random from the tagged version of the Brown Corpus. Forms of *be* and *have* were excluded, as were modal verbs such as *must* and *should.*

> *abandon account acquire act add announce anticipate appear arch ask attempt attend attest avoid bear believe belong bend board boil bring bristle brush build buzz call cap cast choose choreograph close come concern conclude consider contain convert culminate cut deal decrease defend delegate deliver denounce deny depend design determine develop die dine discourage dispatch disunite drink duplicate eliminate emerge end enter equate erect execute exist expect extend face fail fall feed feel fight figure find fly follow get give glow guide hear help hijack hire hope impart impede improve include increase indicate inform instruct inure issue keep learn let live look make mean measure meet mine miss mount mourn near offer open oppose organize own pardon pickle plan play plead prefer prepare present prevent progress project provide question quote range reappear receive recommend remember remind repeat report request resign retire return save say season seat see seem serve set settle shift ship shock sign sing speak spend spice sponsor stand start stay study succeed suffer suggest support surprise swept take talk tell term terminate think touch treat tremble trust try turn understand unite unload use visit weep wheel wipe wish wonder work write*

Appendix B: Complete Output

Of the 193 verbs listed above, Lerner detects 174 in the untagged version of the Brown Corpus. Of these 174, there are 87 for which Lerner does not find sufficient evidence to prove that they have any of the six syntactic frames in question. Some of these genuinely do not appear in the corpus with cues for any of the six, while others do appear with cues, but not often enough to provide reliable evidence. Given more text, sufficient evidence might eventually accumulate for many of these verbs.

The 87 that were detected but not assigned any frames are as follows:

> *account act anticipate arch attend bear bend boil bristle brush buzz cast close contain convert culminate deal decrease delegate deliver depend design determine develop dine discourage dispatch drink emerge end equate erect exist extend fall figure fly glow hire increase instruct issue live look measure mine miss mount mourn open oppose organize own present prevent progress project question quote range reappear receive recommend repeat report retire return season seat settle ship sign sing speak spend sponsor stand stay succeed suffer talk term terminate tremble turn weep wheel*

The 87 verbs for which Lerner does find sufficient evidence to assign one or more frames are shown in Table 10. Reading across each row, a verb is assigned those frames

Table 10
The lexicon that Lerner produces when restricted to the 193 test verbs.

	NP	NPNP	NPcl	NPinf	cl	inf		NP	NPNP	NPcl	NPinf	cl	inf
abandon	NP						make	NP					
acquire	NP						mean					cl	inf
add					cl		meet	NP					
announce					cl		near	NP					
appear					cl	inf	offer	NP					
ask	NP			NPinf			plan						inf
attempt						inf	play	NP					
attest					cl		plead	NP				cl	
avoid	NP						prefer						inf
believe	NP				cl		prepare	NP					
belong			NPcl				provide					cl	
board						inf	remember					cl	inf
build	NP						remind			NPcl			
call	NP	NPNP					request					cl	inf
choose						inf	save	NP					
come						inf	say					cl	
concern	NP						see	NP				cl	
conclude					cl		seem					cl	inf
consider	NP				cl		serve	NP					inf
cut	NP						set						inf
defend	NP						shift	NP					
denounce	NP						shock					cl	
deny	NP				cl		start	NP					inf
eliminate	NP						study	NP					
enter	NP						suggest					cl	
execute	NP						support	NP					
expect	NP			NPinf	cl	inf	surprise					cl	
face	NP						take	NP					
fail						inf	tell	NP	NPNP	NPcl	NPinf		
feel					cl		think					cl	
fight	NP					inf	touch	NP		NPcl			
find	NP				cl		treat	NP					
follow	NP						trust					cl	
get	NP			NPinf		inf	try						inf
give	NP	NPNP					understand	NP					
guide	NP						unload	NP					
hear	NP						use	NP					
help	NP			NPinf		inf	visit	NP					
improve	NP						wipe	NP					
include	NP						wish					cl	inf
inform			NPcl				wonder					cl	
keep	NP						work						inf
learn					cl	inf	write	NP					
let	NP												

whose symbols appear in its row. For easy reference by frame, all the symbols for a given frame are aligned in one column.

Appendix C: Difficult Judgments

The results provided in Tables 6, 7, and 8 are based on hand judgments of the examples found by the cues. In most cases these judgments were clear, but there were five difficult judgments. These five, which were not scored, are discussed below. In all cases except the last Lerner did not find sufficient evidence to warrant a conclusion.

Provide with a tensed clause

Provided and *providing,* when they occur without auxiliaries, clearly take a tensed clause, as in (3a).

3. a. Squat-style lifters and leg-split lifters would both benefit enormously by practicing those variations *providing that they remember* to make alternate sets with the left and right leg to the front.
   b. There must be a restriction in the deed *to provide that the customer may not be charged* more than the current market price for the oil, ...

Provided and *providing* do not appear to be functioning as verbs when they take a tensed clause. Tensed forms of *provide* do not take a tensed clause, as in *"She {provided, provides} that he have enough to live on." However, the infinitive form of *provide* in (3b) is clearly functioning as a verb. The best generalization of these observations is unclear.

Act with an infinitive

(4) "E. B." compared John Brown to Moses in that they were both acting to deliver millions from oppression.

In (4), it is not clear whether the infinitive is an argument, as it would be in "they were both **trying** to deliver ...," or a purpose adjunct, as in "they were both **breaking the law** to deliver ...."

Live with an infinitive

Live occurred twice in the expression *live to see* and once in the expression *live to hear.* If this expression were completely fixed, then its properties should not be considered properties of the verb *live.* However, the infinitive can be any perception verb and possibly a few others. The expression is limited, but not frozen, so the appropriate conclusion is unclear.

Work with a single NP

Work takes a direct object in the sense meaning form or mold. The cues do not detect this sense of *work* in the corpus, but they do find it followed by adverbial NPs such as "every day" and in the expression *work their will.*

Give with a single NP

Lerner judged *give* to take a single NP based on seven examples of which six were mistaken. The seventh was the sentence "They continued to give an arm-elevation." This is ungrammatical in my dialect, but it is clearly an example of *give* with a single NP complement.