

Running head: Structural prediction and ambiguity resolution

Why is *that*? Structural prediction and ambiguity resolution in a very large corpus of English sentences

Douglas Roland, Jeffrey L. Elman, and Victor S. Ferreira

University of California, San Diego

## Abstract

Previous psycholinguistic research has shown that a variety of contextual factors can influence the interpretation of syntactically ambiguous structures, but psycholinguistic experimentation inherently does not allow for the investigation of the role that these factors play in natural (uncontrolled) language use. We use regression modeling in conjunction with data from the British National Corpus to measure the amount and specificity of the information available for disambiguation in natural language use. We examine the Direct Object/Sentential Complement ambiguity and the closely related issue of complementizer use in sentential complements, and find that both ambiguity resolution and complementizer use can be predicted from contextual information.

Why is *that*? Structural prediction and ambiguity resolution in a very large corpus of English sentences

From a certain perspective, linguistic expressions include massive lexical, structural, and acoustic ambiguity. Even when the ambiguities are ultimately resolved by subsequent information (e.g., at the end of a sentence), the incremental nature of most language processing suggests that comprehenders must deal with even temporary ambiguities during the course of sentence comprehension. Yet we seem to understand linguistic expressions with comparative ease, scarcely even noticing any ambiguities at all.

The ways in which comprehenders resolve ambiguity has been a major focus of much research in sentence processing. According to constraint-based accounts of language processing (e.g., Altmann, 1998; Altmann, 1999; MacDonald, Pearlmutter, & Seidenberg, 1994; MacWhinney & Bates, 1989; Spivey & Tanenhaus, 1998), ambiguity is resolved through the interaction of multiple sources of information contained in linguistic expressions. Alternatively, serial accounts of processing (e.g., Frazier, 1978) argue that the initial interpretations of utterances are based solely on syntactic information, and that these interpretations are later revised based on subsequent consideration of other information in the input.

Much of the evidence about the availability and use of information comes from controlled psycholinguistic experiments that typically examine one or two factors at a time. The high degree of control used in experimental designs is essential for determining if and when some theoretically important potential source of information can influence ambiguity resolution.

However, this sort of methodology leaves open a number of questions: Just how ambiguous are linguistic expressions? How much information is available for resolving ambiguity in typical naturally occurring contexts? How large a role do the experimentally studied factors play in natural (uncontrolled) language use and comprehension? Are there other factors that play a significant role in processing? What sorts of interactions occur when a larger number of factors are explored?

Here, we adopt an approach that is complementary to an experimental one. Rather than addressing the question of if and when a particular source of information is used during comprehension, we use corpus data to examine the quantity and variety of information available in normal language use, and to investigate sources of information that would otherwise be missed in the carefully controlled environment of psycholinguistic experiments. Indeed, this approach has the potential to provide information that is relevant to both constraint-based models of sentence processing and syntax-first models of sentence processing. For a syntax-first approach, our results will indicate the degree to which syntax-only heuristics such as minimal attachment can correctly resolve ambiguity, and the extent to which the predictions of the first stage need to be revised by a more general second stage. Our results will also provide information about the information that is used to revise the initial syntax-based predictions. For a constraint-based approach, the analysis of the information available in naturally occurring data provides an indication of the relative importance of factors that are typically investigated in separate experiments, as well as the interaction among these factors.

Knowing how much information is available during normal language comprehension also shows the extent to which the set of factors under consideration can account for language processing. If it turns out that a limited set of factors can be used to resolve nearly all naturally occurring ambiguity, it would suggest that a complex system employing a wide variety of factors may be unnecessarily complicated. On the other hand, if the naturally occurring contexts turn out to be information-poor, it could suggest that an alternate mechanism for ambiguity resolution might be needed.

In order to investigate the amount and variety of information available during natural language comprehension, we identify a relatively large number of general factors that might affect language performance, and investigate those influences using correlational techniques in a very large corpus of naturally produced sentences. We performed three separate analyses on our corpus using regression models.

First, in Study 1, we investigated the amount and type of information available for predicting the resolution of the commonly studied Direct Object / Sentential Complement (DO/SC) ambiguity. The DO/SC ambiguity occurs in sentences fragments such as in example (1), where the post-verbal noun phrase can either be the direct object of the verb, as in (2), or the subject of a sentential complement, as in (3).

- (1) The athlete realized her goals...
- (2) The athlete realized her goals through hard training.
- (3) The athlete realized her goals would be difficult to achieve.

Our main finding of this first set of analyses is that there is abundant information available for resolving the DO/SC ambiguity before the point normally considered to be the disambiguation point - the embedded verb. Given this, we investigate the level of specificity of this information in Study 2. In Study 3, we investigate whether our success in predicting the resolution of the DO/SC ambiguity was in part due to an effort on the part of the producers of the language contained in the corpus data to avoid ambiguity in the first place.

### Study 1: Predicting DO/SC-0 Subcategorization

The goal of the first experiment was to provide a lower bound estimate for the amount of information available for disambiguating ambiguous DO/SC cases during natural sentence processing. This lower bound can alternatively be interpreted as an upper bound on the amount of potential ambiguity during normal comprehension. A secondary goal was to see which types of information are most useful, and which types of information are potentially misleading.

#### *Method*

We began with a large sample of naturally generated language data (described below). The data were coded for a variety of formal and semantic properties, and then used as input to a regression model in order to predict whether each ambiguous example had a direct object continuation or a sentential complement continuation.

The regression analysis used the binary logistic equation shown in Equation 1. Equation 1 allows us to calculate the probability of a given sentence fragment (up to and including the post-verbal NP, but not the subsequent disambiguation region) concluding with either a DO or SC structure (for this equation, values greater than .50 indicate a DO continuation). The equation takes as input a set of predictor variables,  $X_1, \dots, X_p$  (presented below). For each predictor variable, there is a constant,  $B_1, \dots, B_p$ , that represents the weighting of the variable toward a DO

$$\log\left(\frac{p(\text{DO})}{p(\text{SC})}\right) = B_0 + B_1 X_1 + \dots + B_p X_p \quad (1)$$

(or SC) continuation. The constant  $B_0$  (which has no associated predictor variable) represents the a priori likelihood of DOs in the corpus.

### *Corpus Data*

The analysis was carried out using the British National Corpus (BNC) (<http://www.hcu.ox.ac.uk/BNC/>). The BNC consists of approximately 100 million words of written and spoken British English. This corpus was chosen because of its size (smaller corpora do not have sufficient data for the type of analysis in this paper) and balanced nature. Approximately 90% of the BNC is written text, of which three fourths is informative text (applied sciences, arts, belief & thought, commerce & finance, leisure, natural & pure science, social science, world affairs) and one fourth is imaginative text (literary and creative works). The remaining 10% of the corpus consists of transcribed spoken material, including demographic (transcriptions of spontaneous natural conversations) and context-governed material (transcriptions of recordings made at specific types of meetings and events).

For Study 1, a subcorpus was created consisting of the more than 1.2 million sentences in the BNC containing any of the 100 verbs reported by Garnsey, Pearlmutter, Myers, and Lotocky, (1997). These 100 verbs all take both a DO and SC, and many of the verbs in this list have been used by researchers investigating ambiguity resolution. The sentences in this subcorpus were then parsed using the Charniak parser<sup>1</sup> (Charniak, 1997), and automatically labeled for subcategorization. In addition, various key components of the sentences were identified, such as the subject, the post-verbal NP, and the head nouns of these arguments. The search patterns used in this process are described elsewhere (Roland, Dick, & Elman, 2003). A random sample of 100 examples was hand checked for error, revealing that the error rate for the three-way DO/SC/other distinction was approximately 10%.

Because the target of interest in this Study was the DO/SC ambiguity, a smaller set of sentences was extracted that involved only syntactically ambiguous structures. This involved removing all sentences in which the target verb was not used with either a direct object or sentential complement, all sentences in which the sentential complement was marked with the complementizer *that*, and all examples in which the post-verbal NP consisted of an

---

<sup>1</sup> We re-trained the Charniak parser using the (standard) Penn Treebank Brown and Wall Street Journal data in conjunction with additional material created by taking sentential complement examples from these corpora with the complementizer *that* and removing the complementizer in order to improve the parser's performance on complementizer-less examples.

unambiguously case-marked pronoun (e.g., *I, me, he, him, she, her, we, us, they, them*). This left 249,708 examples, or about a sixth of the overall data set, that were ambiguous with respect to DO/SC completion, to the extent that no examples contained the complementizer *that* or case-marked pronouns. This final data set consisted of 181,692 (72.8%) DO examples and 68,016 (27.2%) SC examples.

### *Regression Model Input*

Four types of information, which will be described in detail below, were used as predictors in the regression model: verb identity or *lemma* information, length, frequency, and semantic information. For each sentence, the value of each variable was encoded, as well as the outcome variable (DO or SC). Within each analysis, all predictors were added into the regression in a single step.

*Verb lemma.* Different verbs have different subcategorization biases. The verb preceding an ambiguous NP has been shown to play an important role in how the NP is interpreted (e.g., Garnsey et al., 1997; Trueswell, Tanenhaus, & Kello, 1993). If the verb more commonly occurs with direct objects, then comprehenders are more likely to interpret the ambiguous post-verbal NP as a direct object, whereas if the verb more commonly occurs with a sentential complement, comprehenders are more likely to interpret the NP as the subject of the complement. Verb bias is captured in the model by including verb lemma as a categorical variable. The verb lemma is a single representation for the verb across all inflected forms. Thus, examples containing the verb word forms *feel, feels, feeling, and felt* would all be identified as containing the lemma *feel*.

*Length.* We use several length measures as predictors in the regression model. The length of the ambiguous post-verbal NP has been shown to affect the resolution of DO/SC ambiguities (F. Ferreira & Henderson, 1991). Additionally, length has been found to play a role in determining whether a complementizer is used in sentential complements (e.g., Hawkins, 2002). Additionally, length is highly correlated with the pronominal or full NP status of a given NP.

To incorporate length into our model, we included the following specific measures of length: the length (in characters, including spaces<sup>2</sup>) of the subject NP, the length of the head noun of the subject NP, the length of the post-verbal NP, the length of the first word of the post-verbal NP, and the length of the head noun of the post-verbal NP. Thus, a sentence such as *The old man felt the restaurant bill was exorbitant* would have the following five length values associated with it: subject NP length = 11 (*The old man*), subject NP head = 3 (*man*), post-verbal NP = 19 (*the restaurant bill*), post-verbal NP first word = 3 (*the*), and post-verbal NP head = 4 (*bill*).

*Frequency.* Although lexical frequency has been shown to play a role in many areas of language processing, it has not been shown to play a role in DO/SC disambiguation (although it may be relevant to whether the complementizer *that* is expected in sentential complement

---

<sup>2</sup> The database used to handle data had a maximum possible field length was 255 characters. This resulted in all NPs being truncated at 255 characters, and all measures being based on the first 255 characters. Thus, the maximum possible length for the subject NP, post-verbal NP, etc. was 255 characters. Approximately 1.6% of all post-verbal arguments in the BNC have a length of 255 characters or greater. We measured length in characters (as opposed to potentially more psycholinguistically plausible measures such as length in syllables or words) purely for practical reasons.

structures - see Juliano & Tanenhaus, 1993). However, like length, it is possible that frequency may be correlated with other factors, such as pronoun versus non-pronoun, function word versus content word, and so forth, and these may play a role in DO/SC disambiguation. We included three frequency measures as predictors in the regression model. These are the logarithm-transformed frequencies<sup>3</sup> of the subject NP head noun (e.g., *man* in the above example), the post-verbal NP first word (e.g., *the* in the above example), and the post-verbal NP head noun (e.g., *bill* in the above example). Log frequencies were generated by taking the logarithm of the word count of the word form, ignoring capitalization, from the BNC. Thus, *cat* and *cats* would have different frequencies, while *cat* and *Cat* would have the same frequency.

*Semantics.* A number of semantic factors have been shown to play a role in ambiguity resolution. For example, Garnsey et al. (1997) showed that the plausibility of the post-verbal NP as a direct object for the main verb plays a role in DO/SC ambiguity resolution, and McRae, Spivey-Knowlton, and Tanenhaus (1998) demonstrated that the thematic fit of the NP (as either an agent or patient) plays a role in ambiguity resolution in reduced relative clauses.

Typically, plausibility and thematic fit measures rely on human judgment. Because we could not get human judgments for the large number of sentences that we assessed, we used automatically generated semantic representations for the subject NP, the subject NP head noun, the post-verbal NP, and the post-verbal NP head noun. These representations were based on Latent Semantic Analysis (LSA) (Deerwester, Dumais, Furnas, Landauer, & Harshman, 1990). LSA relies on a word's pattern of co-occurrence statistics as a proxy for its meaning. Singular Value Decomposition is used to perform dimension reduction on the word co-occurrence matrix and condense the information into a high-dimension semantic space. Any word or phrase can then be represented as a vector in the resulting semantic space. The cosine of the vectors representing two words or phrases has been shown to correspond with human judgments of semantic similarity for the same words or phrases in a wide variety of tasks (e.g., Kintsch, 2001; Thomas. K. Landauer & Dumais, 1997; T. K. Landauer, Foltz, & Laham, 1998; T. K. Landauer, Laham, & Foltz, 1998; T. K. Landauer, Laham, Rehder, & Schreiner, 1997; Rehder et al., 1998; Wolfe et al., 1998).

We generated a semantic space based on all of the texts occurring in the BNC using the LSI (LSA) tools from Bellcore/Telecordia. We then found the locations for each subject NP, subject NP head noun, post-verbal NP, and post-verbal NP head noun in this semantic space<sup>4</sup>. All vectors were normalized to a length of 1 based on the first 400 dimensions, because the direction of the vector reflects semantics, whereas the length reflects the frequency of the word or words in the co-occurrence matrix. We then used the values for the 20 most important dimensions/semantic factors, based on the amount of variance from the original co-occurrence matrix that each factor captured, as our representation for each item. In this manner, the subject NP, subject NP head noun, post-verbal NP, and post-verbal NP head noun are each represented by a 20 unit vector. Because LSA typically works best with the first ~300 dimensions, we are losing potentially useful information by only considering the first 20 dimensions. However, the

---

<sup>3</sup> Base 10.

<sup>4</sup> We do not provide the model with a separate semantic representation of the verb (or verb length and frequency information), since any such representation would contain less information than is already provided through the *verb lemma* categorical variable.

benefit is that (a) it limits the ability of the regression model to find coincidental patterns in the data, and (b) this allows the use of a variety of commercial software (e.g., SPSS, Microsoft Access) which has a limit of 255 analysis variables.

### *Results*

We performed several different analyses using the data and regression model described above. The first set of results reflects the overall performance of the model when all information is entered in to the regression in a single step. This measures the combined contribution of all sources of information. The second set of results reflects the performance of the model when only specific subsets of the data are used. This measures the amount of information contributed by these separate sources of information. The third set of results reflects the performance of the model when various sets of data are added as a second step to a model that has already been given a certain set of information. This measures the additional information contributed by the second source of data above and beyond that contributed by the first source of information.

#### *Overall Contribution of all Sources of Information*

The overall results of the regression model are shown in Table 1. A minimum or baseline of performance for this model can be determined by simply having the model always guess the most frequent structure under consideration, which in this case is DO. If the model always predicts a DO structure, a baseline of 72.8% results. When instead the model uses all the above information to guess structure, it was able to correctly predict the DO/SC continuation of 86.4% of the examples. There is a significant improvement in model fit over the baseline model (Nagelkerke R square = 0.568, Model Chi Square (186, N = 249708) = 124311.496,  $p < 0.001$ ).

	Baseline %	Model %	Nagelkerke R square	Model Chi Square
Main model (all info)	72.8%	86.4%	0.568	$\chi^2$ (186, N = 249708) = 124311, $p < 0.001$
Lemma only	72.8%	83.1%	0.463	$\chi^2$ (98, N = 249708) = 96143, $p < 0.001$
Semantic only	72.8%	79.8%	0.331	$\chi^2$ (80, N = 249708) = 64785, $p < 0.001$
Length only	72.8%	77.1%	0.241	$\chi^2$ (5, N = 249708) = 45441, $p < 0.001$
Frequency only	72.8%	77.7%	0.190	$\chi^2$ (3, N = 249708) = 35001, $p < 0.001$
Subject info only	72.8%	72.8%	0.137	$\chi^2$ (43, N = 249708) = 24815, $p < 0.001$
Post-verbal NP info only	72.8%	79.5%	0.325	$\chi^2$ (45, N = 249708) = 63377, $p < 0.001$

Table 1. Binary Logistic Regression Models for predicting SC-0 versus DO.

In an additional test to investigate the ability of the model to generalize, a simulation was run in which a random 10% of the corpus data was withheld before the values of the regression model were set. When subsequently tested on this unseen 10% of the data, the model performed at nearly the same level as the model trained and tested on the full set of corpus examples (86.6% vs. 86.4%). This indicates that the final performance of the model does not reflect an over-fitting of the data, but reflects the discovery of reliable relationships between the predictor variables and the predicted structures.

#### *Contributions of Various Subsets of Information*

We evaluated the performance of the regression model when given only certain subsets of input information in order to evaluate (a) how much information is available at different points during sentence processing, and (b) how much information is available from different types of information (frequency, length, semantic, lemma). For example, to determine the amount of information available from the grammatical subject only, we used (only) the frequency, length, and semantic vectors associated with the subject as predictors in the regression model. As with the complete regression model, baseline performance is 72.8%. The contributions of each source of information are shown in Table 1.

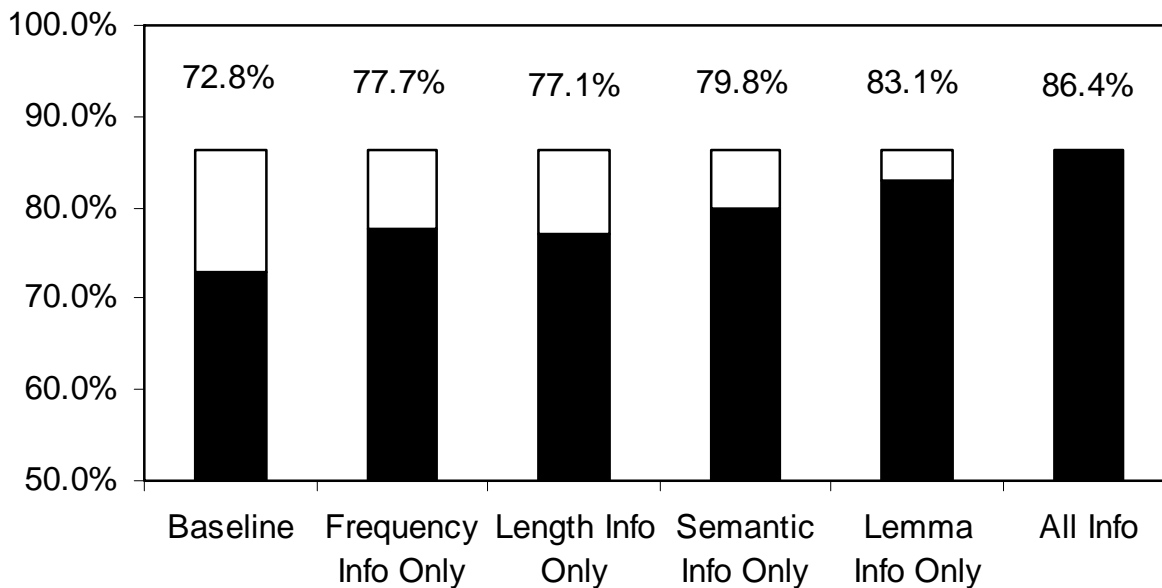


Figure 1. Performance of DO/SC regression model with information type based subsets of information.

The contribution of these different types of information can be organized in two ways. One is to organize information sources as to the type of information: lemma identity, semantic information, length, and frequency information (see Figure 1). These analyses reveal that the gains in model performance when different kinds of information are added vary from an additional 4.9% correct (frequency only) to an additional 10.3% correct (lemma information only). Each of the sources of information provides a significant improvement in model fit and an

improvement in the performance of the model. (Note that gains in performance when the model is already at a higher level of performance are more important than comparable numeric gains when the model is at a lower level of performance.)

Alternatively, information sources can be organized in terms of when in the sentence they can contribute to comprehension when processing from beginning to end, by considering the contributions of the factors associated with the main subject, the lemma, and the post-verbal NP respectively. These data are shown in Figure 2. Here, adding main-subject NP information only results in a gain in performance of less than an additional 0.1% correct, adding lemma information (again) adds 10.3%, and adding post-verbal NP information adds 6.7%. The information associated with each of the regions results in a significant improvement in model fit, and each of the regions except the subject region also results in an improvement in the performance of the model.

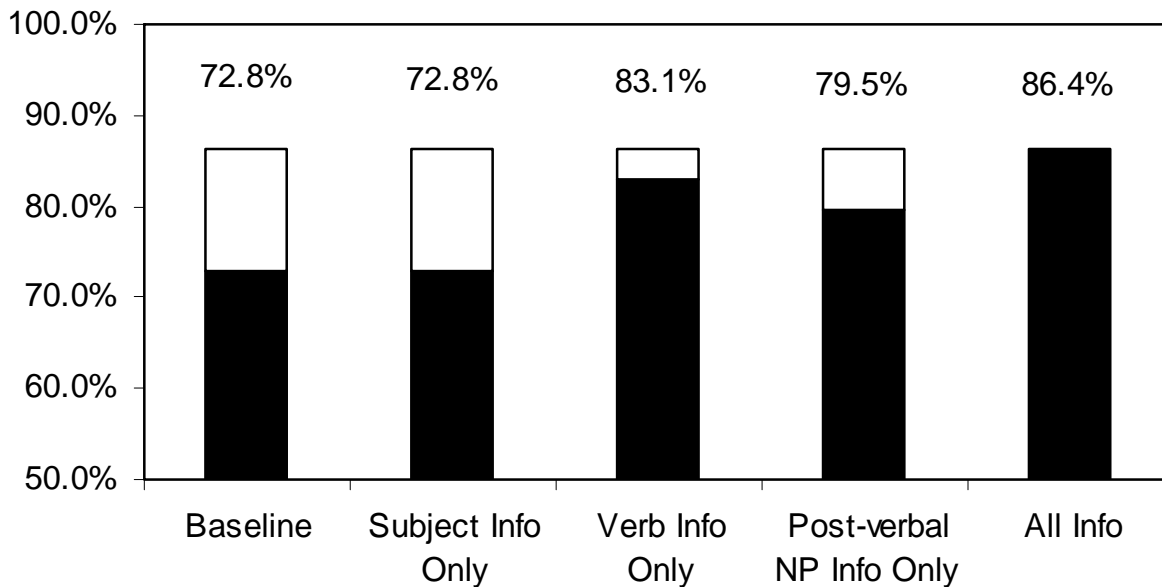


Figure 2. Performance of DO/SC regression model with location based subsets of information.

In the above analyses, the information associated with knowing the verb lemma makes the largest contribution to the overall results of the model. However, the knowledge gained by knowing the verb lemma overlaps with the knowledge gained from the other sources of information (e.g., knowing that you have a particular lemma may cause such a strong bias towards concluding that the post-verbal NP is a direct object, that as a result, knowing that the post-verbal NP is long may not provide any additional bias). In order to measure the unique contributions of each of the sources beyond the information gained from the verb lemma, we ran several two-step regression models. In these models, the lemma information was added in the first step, essentially increasing the baseline from 72.8% to 83.1%. Then, different other sources of information were added in a second step, and the resulting gain in performance of the model was measured. The results are shown in Table 2.

*Additional Contributions of Various Sources of Information when Added to Lemma Information*

	Baseline % (includes Lemma)	Model %	Nagelkerke R square	Step (for information to lemma data)	Chi adding to model	Square additional with
All Information	83.1%	86.4%	.568	$\chi^2$ (88, N = 249708) = 28167, p<0.001		
Semantic only	83.1%	86.1%	.559	$\chi^2$ (80, N = 249708) = 25461, p<0.001		
Length only	83.1%	84.5%	.517	$\chi^2$ (5, N = 249708) = 14126, p<0.001		
Frequency only	83.1%	83.7%	.499	$\chi^2$ (3, N = 249708) = 9254, p<0.001		
Subject info only	83.1%	83.1%	.471	$\chi^2$ (43, N = 249708) = 2042, p<0.001		
Post-verbal NP info only	83.1%	86.2%	.566	$\chi^2$ (45, N = 249708) = 27512, p<0.001		

Table 2. Binary Logistic Regression Models Adding information to Lemma only model for predicting SC-0 vs. DO.

As above, these results cover both the information-type based subsets of semantic information, length information, and frequency information, as well as the location-based subsets of information including main subject and post-verbal NP. Results are comparable to the analyses performed without adding lemma information first, with additional variance accounted for varying between less than 0.1% (main subject NP information only) to 3.0% (semantic information only). All subsets of information result in a significant improvement in model fit, and all subsets except the subset associated with the subject region result in an improvement in model performance.

*Performance of Separate Models for Each Verb*

In the above analyses, the regression model must choose a single coefficient (B) for each of the predictors. This imposes a limitation on the model, in that it must optimize these values for use across all verb lemmas (i.e., the model cannot take into account interactions among the predictor variables). In order to (a) provide more detailed information about how much information was gained on a verb-by-verb basis from sources beyond the identity of the verb, and (b) to eliminate the inherent constraints of the regression model against setting separate weightings for each of the predictors for each verb, a separate sub-analysis was performed. In this sub-analysis, 100 separate versions of the regression model were run, one for each verb. With the obvious exception of the *lemma* categorical variable, all input was otherwise the same as in the main regression model. The results of this sub-experiment are shown in Figure 3. Accuracy in prediction is shown for each verb, along with a verb-specific baseline. The previous baseline of 72.8% is no longer relevant, since each verb has its own DO/SC bias.

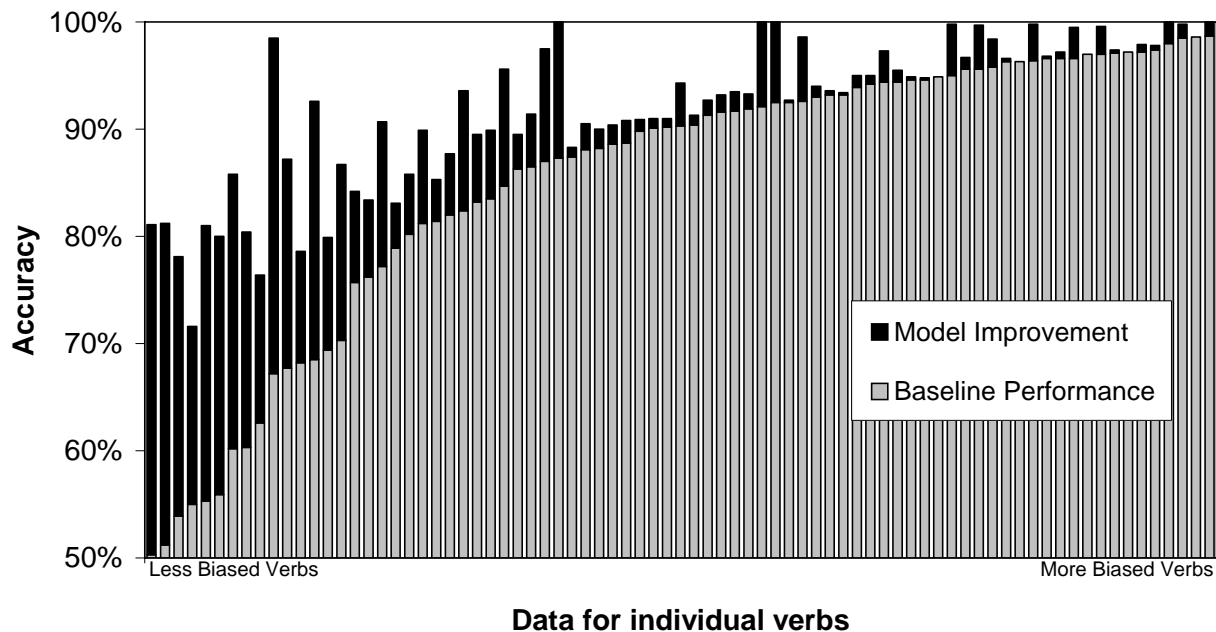


Figure 3. Performance of separate regression models for each verb.

The addition of length, frequency, and semantic information improved the performance of the regression model for a large majority of the verbs, as can be seen in Figure 3. The improvement in model fit over the baseline model was significant for 72 verbs, and non-significant for 7 verbs (the regression model was unable to converge on a solution for the remaining 21 verbs when the default parameters were used<sup>5</sup>). However, Figure 3 also reveals that there is considerable variation in the ability of the model to predict the correct subcategorization, depending on the verb. Obviously, verbs that have an extremely strong bias, such as *conceal* or *advocate* (which are almost always used with DOs) provide little opportunity for improvement above their baseline predictions. All 7 of the verbs where the model shows no significant improvement in fit over the baseline model have a baseline starting point above 84%, and 3 of these verbs have a baseline above 98%. An alternate way of looking at the performance of the model on these 7 verbs is to consider the Nagelkerke R square of the model for each of these verbs, which is above .85 for all but one of the seven verbs. This suggests that for these six verbs, the lack of significance in the models' performance is due to a high baseline. The remaining case, *observe*, remains unexplained (*observe* has a baseline of 94.6%, and a model performance of 94.8%, Model Chi Square (88, N = 996) = 103.714,  $p = 0.121$ ).

It is also useful to look at the collective performance of the separate verb-by-verb models. We provide frequency-weighted results to allow for direct comparison with the cross-verb model in the previous section.

<sup>5</sup> With relaxed convergence parameters, the model will converge on a solution for 94 out of 100 of the verbs. The problem appears to be related to co-linearity of some predictors within the smaller sample sizes of a single verb.

The average baseline for the separate models, weighted by verb frequency, is 83.0%. Note that this is the same value<sup>6</sup> as the performance of the cross-verb model when the lemma information is taken into account. The mean performance, averaged over all verb-specific models weighted by verb frequency, was 88.3%. The difference between the final performance of 88.3% for the separate models and the 86.4% performance of the single model in the previous section indicates the degree of improvement which can be obtained by allowing the model to set separate regression coefficients for each lemma (i.e., interaction between verb information and the other information sources), rather than having to set a single value that is optimized across all lemmas.

### *Discussion*

#### *Overall Model Performance*

Overall, the model when given full access to all sets of predictors achieved an 86.4% level of accuracy. In order to interpret this level of accuracy, one must consider what reasonable lower and upper bounds of performance of the model might be.

*Lower bound.* Although the goal of the model is to make a binary DO/SC decision, the most appropriate baseline is not 50%, but rather 72.8%. This is because the data set contains 72.8% DO examples, so that any model that always guessed DO (such as a Minimal Attachment model - e.g., Frazier, 1978) would be guaranteed to get 72.8% of the predictions correct by default. Any performance in excess of this baseline is evidence that the model has found a correlation between (at least) one of the input variables and the actual structure.

*Upper bound.* Hand analysis of a random subset of the sentences revealed an error rate of approximately 10% in the DO versus SC categorization of the input data. One way of looking at the problem this introduces is that if the model somehow correctly predicted the true continuation of 100% of the examples, 10% of these predictions would really be incorrect, because 10% of the data were incorrectly labeled by the automatic parser with a wrong subcategorization. A second way of looking at this is that the noise induced in the model by this mislabeled data will presumably limit the ability of the model to make correct predictions. In either case, the 10% error rate can be expected to impose an upper bound of a 90% level of accuracy in predicting sentence continuations.

Consequently, the model's overall level of performance of 86.4% is best compared to a highest possible level of performance of 90%. This suggests that in terms of information availability, the overall degree of actual ambiguity in natural language use is low, even in a set of examples chosen due to their potential for ambiguity. Out of more than 1.2 million examples of the 100 SC-taking verbs assessed in these regressions, approximately 250,000 examples were syntactically ambiguous (no complementizer, no unambiguously case marked pronouns), and of these, the model was able to successfully predict the continuation of 86.4% of the examples.

---

<sup>6</sup> The difference between 83.0% for the average of the separate models and 83.1% for the cross-verb model with lemma information is due to the cross-verb model containing data for all 100 verbs, while the average of the separate models contains data for only the 77 verbs where the separate models converged on a result.

Furthermore, for a variety of reasons, this result can be viewed as an underestimate of the amount of information available to a general model like the one assessed here. First, the information available to the model under-represents what is actually available to humans during sentence processing. For example, when the model sees a pronoun (represented in this analysis only as a short, high frequency item with a particular semantic representation), the model does not have any access to information about the referent of the pronoun, whereas a human subject presumably does have access to such information.

In addition, the logistic regression model has limitations. In particular, the model assumes that sentence continuation probabilities are linearly related to each of the predictor variables, but this may in fact not be the case. Furthermore, the model does not allow for interactions between the variables. One specific example of this is that the model cannot set separate values for each parameter for each lemma (except inasmuch as tested by the verb-specific models tested above). The model can only set a cross-verb bias for each parameter, and then a separate DO/SC bias for each verb.

Finally, as indicated above, the 10% error rate in the labeling of the subcategorizations in our dataset, as well as an additional degree of error induced from errors such as the mislabeling of head nouns, etc., makes it unlikely that the model can perform any higher than about 90%. This simulation then suggests a profound conclusion: In principle, linguistic expressions -- even ones chosen for their potential to be ambiguous -- might in fact be almost completely unambiguous.

### *Contributions of Individual Factors*

The largest contribution to the overall performance of the model came from the information associated with the verb lemma. In effect, this is because knowing what the verb lemma is allows the model (or the comprehender) to switch from an overall bias towards a DO interpretation (minimal attachment) to having a different bias in the case where an individual verb actually has a SC bias. All of our cross verb models in the above experiments are essentially limited to placing a metaphorical thumb on the scale to shift the bias towards DO or SC depending on the lemma (via the weightings chosen for the lemma categorical variable). However, humans and the final set of models described above, where separate sets of regression parameters are found for each verb, can do additional fine tuning based on the lemma, and make decisions along the line of “usually a long post-verbal NP indicates a DO, but for *this* verb, it suggests an SC”. In general, this finding of the significance of the identity of the verb fits in with experimental results such as Trueswell et al. (1993) and Garnsey et al. (1997), who showed that the subcategorization bias of the verb affects the parsing difficulties in sentences with the DO/SC ambiguity. If the verb is more frequently used with a direct object (i.e., has a DO bias), then parsing is more difficult in the region after the post-verbal NP than it is if the verb is more frequently used with a sentential complement (i.e., has a SC bias).

Our model also shows that the post-verbal NP is important in resolving the DO/SC ambiguity. To the extent that semantic information contributes to the information contained in the post-verbal NP, this performance reflects the experimental results in Garnsey et al. (1997), who showed that both verb bias and the plausibility of the noun phrase following the verb as a direct object played a role in the interpretation of DO/SC examples. The individual LSA-based

semantic factors that we used cannot be directly interpreted with respect to notions such as thematic fit of the NP for a specific verb, or even with respect to more general notions such as animacy or other common semantic notions. However, we did sort the post-verbal NPs and post-verbal NP heads by their predicted strengths as cues for either DO or SC interpretations, and assigned post hoc interpretations to the type of word or NP that the model preferred for each continuation. The general tendency was for SC-biased items to be animate, and for DO-biased items to be inanimate or at least more abstract entities than those found as SC subjects.

The strength of information associated with the post-verbal NP is not just due to the semantic information contained in the post-verbal NP. Length also played an important role. This appears to be the result of several factors. Pronouns, which occur more frequently with SC completions, are short (1-3 characters), while full NPs typically are much longer. Additionally, object NPs are much more likely to be modified than (embedded) subject NPs, which also contributes to the length factor. The most important of all of the length variables in the model was the length of the post-verbal NP, with long post-verbal NPs providing evidence for a DO interpretation.

The role of the pronominal versus full NP status of the post-verbal NP in resolving the DO/SC ambiguity was also reflected in the frequency measures. The most important of the frequency variables was the frequency of the post-verbal NP head noun, with high frequency heads (e.g., pronouns) favoring a SC interpretation. Note that in the case where the head is a pronoun, this same pronoun is also the first word of the post-verbal NP. However, the effect of high-frequency pronouns in the first word position was counterbalanced by the presence also of the high frequency determiner *the* in the same position. *The* occurs much more frequently in direct objects than in sentential complements.

Within our model, the information associated with the main subject plays a much smaller role in resolving the DO/SC ambiguity. However, the subject presumably plays a major role in other aspects of processing the sentence, such as in the initial estimates, for example, of whether the sentence is likely to be a sentence complement or transitive structure at all.

### *Verb Specific Models*

One of the sets of analyses performed in this section of the paper involved a series of separate models for each of the verbs. These results are shown in Figure 3. The main interpretation of these results is that contextual information is available for resolving the DO/SC ambiguity for most of the verbs. Most of the overall gain in performance occurs in the verbs on the left-hand side of Figure 3 – the verbs with the lowest baselines (i.e., they have approximately equal probabilities of a DO or SC completion), and therefore the most potential for improvement.

However, the unsurprising nature of this observation hides an important issue. These verbs have been of interest to psycholinguists, because it is assumed that they maximize the possibility for ambiguity, insofar as their typical pattern of usage does not favor one or the other structure. However, the ability of the regression model to predict how the ambiguity will be resolved in the majority of these seemingly equi-biased verbs suggests that the additional information provided by the subject and post-verbal NP is highly informative of what the actual structure will be. This suggests that although these verbs may be equi-biased in a generic sense

(when the specifics of the usage are ignored), the verbs may have a fairly strong bias towards a particular completion in a specific context. This is consistent with Hare et al.'s (2003) claim that the subcategorization usage of many equi-biased verbs is correlated with different senses of the verb; one would expect that some of our variables are correlated with different verb senses, and thus provide an indirect indication of the likely subcategorization.

### *What Does the Model Get Wrong?*

In examining the model's performance, we found that having a pronominal, short, or animate post-verbal NP favored an SC interpretation, whereas having a full, long, inanimate, or abstract post-verbal NP favored a DO interpretation. Although the model is able to correctly predict the continuation of most of the examples, the model was not able to predict the correct continuation (i.e., the model "garden pathed") of 13.6% of the ambiguous examples.

To some extent, this is the result of the limitations discussed above, but there are also cases where the model "legitimately" makes mistakes. These cases tended to be examples where the corpus sentence violated the *short/animate/pronoun suggests SC* and *long/full/abstract/inanimate suggests DO* patterns. Examples of this include (4) and (5), where the model misinterpreted the highlighted NPs as being direct objects rather than subjects of the complements (shown in parenthesis). Another example is shown in (6), where the model interpreted *there* as being the subject of a sentential complement. For reference, the original context is shown in (7).

- (4) I found **the answer** (was a machine called a copy lathe).
- (5) I understand **the Campaign for Real Ale Limited** (may change the amount only after giving me prior notice).
- (6) I think **there**.
- (7) Outside, that was some stabling. I think there.

### *Conclusions for Study 1*

The data in Study 1 show that the DO/SC structural ambiguity, although not rare, only occurs in approximately one sixth of naturally occurring examples containing the 100 verbs that are most likely to have this ambiguity. Additionally, the data show that a structural heuristic such as minimal attachment would allow the comprehender to correctly predict the resolution of 72.8% of the structurally ambiguous examples. Our regression model provides evidence that the use of additional contextual information would allow for the correct resolution of at least 86.4% of the structurally ambiguous cases. This suggests a low overall rate of remaining ambiguity. Whether or not comprehenders use all of the available information is beyond the scope of the present study, and can only be determined experimentally. Additionally, the results suggest that the identity of the verb (lemma) and the nature of the post-verbal NP play a large role in DO/SC disambiguation, while the main subject of the sentence plays a lesser role.

### Study 2: Predicting DO/SC-*that* Subcategorization With a DO/SC-0 Trained Model

One of the findings of the first study is that the nature of the temporarily ambiguous post-verbal NP contributes a significant amount of information towards the resolution of the

ambiguity – in other words, the ambiguous NP itself has a lot to say about whether it is more likely to be a direct object or a subject of an embedded clause. However, this raises a second question: What granularity of representation is needed to correctly distinguish between NPs which are likely to be objects and those which are likely to be subjects? Is it sufficient to be able to distinguish between prototypical subjects and prototypical objects, or must the comprehender make finer distinctions between different types of subjects and objects?

We examine this issue using data from a structure related to the ambiguous sentential complement (8) that formed one part of the DO/SC ambiguity, the unambiguous sentential complement with the complementizer *that*, shown in (9).

(8) The athlete realized her goals would be difficult to achieve.

(9) The athlete realized that her goals would be difficult to achieve.

We prepared a second data set consisting of pseudo-ambiguous examples made by removing the complementizer *that* from the unambiguous examples, to create a set of examples where the embedded subjects should be maximally similar<sup>7</sup> to those in the ambiguous sentential complements. If the relevant information for resolving the DO/SC ambiguity is the ability to distinguish between the types of items that typically serve as subjects and those that typically serve as objects, then the embedded subjects of cases such as (8) and (9) will presumably be quite similar, whereas if we find that the embedded-subject NPs of (8) and (9) differ greatly in their properties, then this would suggest that a simple representation of the differences between typical subjects and typical objects is not sufficient for resolving the DO/SC ambiguity.

The regression model developed in the first study allows us to examine whether there is a broad-based difference in natural language use between *SC-that* and SC-0 examples. The model in the previous section was trained to distinguish between the contextual information accompanying DO and SC-0 examples. If SC-0 and *SC-that* contexts are equivalent, then the SC-0 trained model should be able to distinguish just as well between DO and *SC-that* pseudo-ambiguous examples as it did between DO and SC-0 cases, based on the same type of contextual data. This would suggest that a fairly broad representation of subject versus object is sufficient for DO/SC ambiguity resolution.

Alternatively, if the SC-0 trained model cannot distinguish between DO and *SC-that* examples, then it suggests that SC-0 and *SC-that* contexts are not equivalent, and that the representations necessary for DO/SC ambiguity resolution are more specific than a broad representation of subject versus object.

### *Methodology*

This study used the regression model developed in the previous section. Thus, the training data (on which the regression model's parameters are set) consist of the same 249,708 DO and SC-0 corpus examples used in the first study. These examples were coded for the same

---

<sup>7</sup> Since our null hypothesis is that all subjects should have similar properties, we are being conservative by comparing only subjects which appear in very similar structures.

length, frequency, semantic, and lemma properties used in the first study. Additionally, a second test set of data was prepared. This second set of data consists of the 74,940 BNC corpus examples where the 100 target verbs are used with *SC-that* sentence completions. As with the DO and SC-0 examples, all sentences in which the post-verbal NP was an unambiguously case-marked pronoun were removed. Since the complementizer obviously marks the remaining sentences as SCs, it also was deleted from all examples before preparing the length, frequency, and semantic information for use in the regression model. The regression model was first run with the 249,708 example DO/SC-0 training data set, with all predictors being added in a single step. Once the regression model parameters were set, the model was tested with the second set of pseudo-ambiguous *SC-that* data, and the model's DO/SC predictions for this set were recorded.

### *Results and Discussion*

We performed three separate analyses on the data generated by testing the pseudo-ambiguous *SC-that* examples on the DO/SC-0 trained regression model. In the first analysis, we examine the subcategorization predictions for the pseudo-ambiguous examples. Recall that on the DO/SC-0 data, the model began with a 72.8% baseline, and results in an overall performance of 86.4% on the DO/SC prediction task. However, in order to fairly evaluate the performance of the model on the *SC-that* data, we must break down the performance of the model on the original data into performance predicting the completion of DO examples and the performance predicting the completion of SC-0 examples. The overall performance level of 86.4% breaks down into a 92.5% success rate in predicting the continuation of (the more frequent) DO examples, and a 70.1% success rate for the SC-0 examples<sup>8</sup>.

This means that if the *SC-that* examples are similar to the SC-0 examples, we would also expect approximately a 70.1% level of performance in sentence completion prediction on the *SC-that* examples. However, as Figure 4 shows, the level of performance in this task is only 24.6%. This shows that a model trained to distinguish SC-0 and DO structures is unable to distinguish pseudo-*SC-that* structures from DO structures. This indicates that *SC-that* and SC-0 examples do not have similar formal and semantic properties, but instead that the properties of the *SC-that* examples are more similar to those of the DO examples than to those of the SC-0 examples.

Our second analysis of this data set investigated the dimensions along which the *SC-that* examples and DO examples are similar. In order to do this, we examined the average values for each class (DO/SC-0/*SC-that*) of each of the 88 length, frequency, and semantic factors that were used as predictors in training and testing the regression model. Thus, for each of the 88 predictors, we calculated the average value for that predictor across the 181,692 DO examples, the average value for the 68,016 SC-0 examples, and the average value for the 74,940 *SC-that* examples. If the SC-0 and *SC-that* shared similar properties, one would expect the average values of their predictors to be similar, whereas if the predictors were unrelated to subcategorization at all (i.e., random), one would expect by random chance the values for *SC-that* to be more similar to those of SC-0 half of the time, and more similar to those of DO the other half. Instead, what we find is that for 69 of 88 (78.4%) of the predictors, the average value of the *SC-that* examples was closer to the average value for the DO examples than to the average

---

<sup>8</sup> 92.5% x 181,692 DO examples + 70.1% x 68,016 SC-0 examples = 86.4% overall performance.

value for the SC-0 examples. This shows that not only are the SC-0 and SC-that examples different from each other, but that the SC-*that* examples are in fact similar to the DO examples.

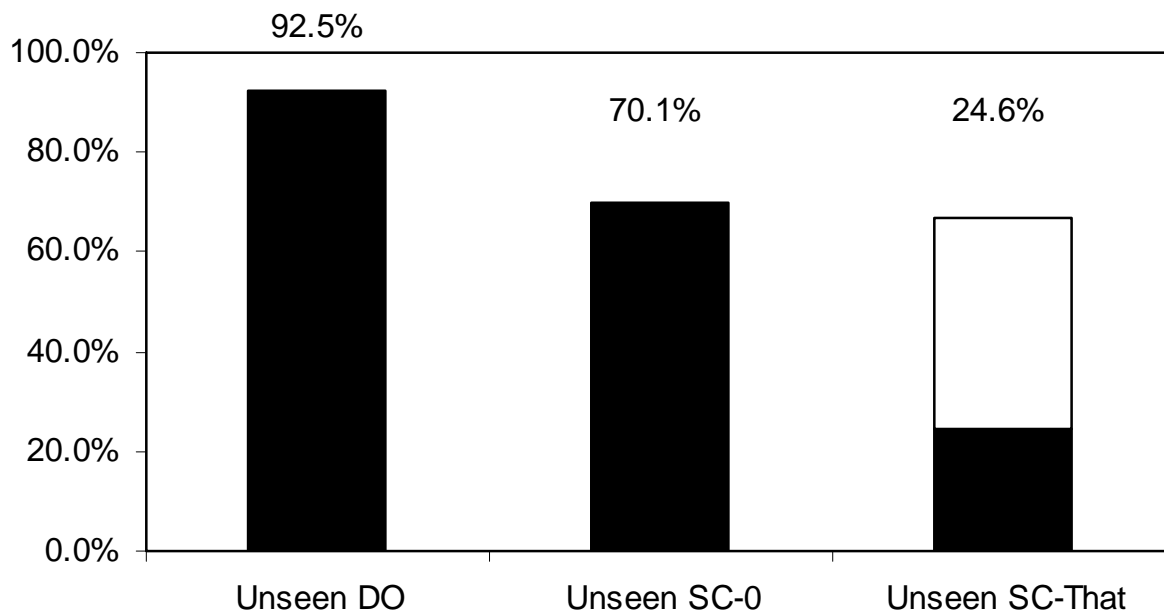


Figure 4. Performance of regression model on SC-that data.

Although for most predictors, the SC-*that* and SC-0 examples did not have similar values, it is the case that for some predictors, the SC-*that* values were more like the SC-0 values than they were like the DO values. The factors where SC-*that* and SC-0 were more similar were nearly all<sup>9</sup> related to the main clause subject rather than the post-verbal NP. This indicates that although the subjects of SC-0 and SC-*that* sentences tend to be similar, the post-verbal NPs of DO and SC-*that* examples are more similar to each other than they are to the post-verbal NPs of SC-0 examples. This is consistent with a notion that the SC and DO uses of verbs are different senses, with corresponding differences in the subjects (cf., Hare et al., 2003) whereas the complementizer *that* is more likely to be used in the SC cases where the post-verbal NP is similar to the post-verbal NPs that typically occur with DO examples.

Our third analysis of this set of data investigated a possible confounding factor<sup>10</sup> in our results for Study 2. It is possible that the difference in performance of the model between the SC-0 and SC-*that* examples might arise even if the length, frequency, and semantic properties of the SC-0 and SC-*that* examples were identical, but the distribution of verbs in the SC-0 and SC-*that* data sets was different. An apparent decrease in performance on the SC-*that* data relative to the SC-0 data could then result if the model tended to make better predictions on verbs which occurred more frequently without the complementizer *that* than with the complementizer.

<sup>9</sup> These were subject head noun frequency, subject NP length, and 16 subject-related semantic factors vs. only 5 semantic factors related to the post-verbal NP.

<sup>10</sup> We thank Doug Rohde for this suggesting this possibility.

In order to test this hypothesis and to eliminate the possible effects of verb distribution on our results, we re-ran the regression model on a verb-by-verb basis (i.e., a separate model for each verb, using the DO and SC-0 data for training, and the *SC-that* data for testing). The initial portion of this experiment – running separate DO/SC-0 models for each verb – is the same as shown in Figure 3. The additional element in this analysis is to evaluate the performance of these models on the pseudo-ambiguous *SC-that* data.

The results are summarized in Table 3. We analyzed these results in two separate ways. The first, using raw data, provides for a comparison of the separate models with the performance of the single cross-verb model shown in Figure 3. The second, using normalized data, eliminates the possible effects of verb distribution.

	Average Performance (% Correct out of all examples)	Average Performance (Normalized)
DO Examples	93.3%	94.1%
SC-0 Examples	74.8%	50.9%
<i>SC-that</i> Examples	35.8%	25.9%

Table 3. Cross verb summary of individual models regression models for each verb, trained on DO and SC-0 data, tested on *SC-that* data.

As in the case of the data shown in Figure 3, the regression models converged for only 77/100 of the verbs. The first column in Table 3 shows the results for the raw data. As in the case of the single cross-verb model shown in Figure 3, the overall performance was 88.3%, over a baseline of 83.0%. The overall result of 88.3% breaks down into a 93.3% performance on DO examples, and a 74.8% performance on SC-0 examples. As in cross verb results shown in Figure 4, the performance of the separate DO/SC-0 trained models on the *SC-that* examples was much lower (35.8%) than the performance on the SC-0 examples.

The second column in Table 3 shows the average model performance when we normalize the results so that the results of each of the 77 different verb/models have the same weight. These results show that the performance on SC-0 examples is still higher than the performance on *SC-that* examples, indicating that there are in fact intrinsic differences between the SC-0 and *SC-that* examples with respect to the length, frequency, and semantic measures.

### *Conclusions for Study 2*

The combined results of the three analyses in Study 2 demonstrate that *SC-that* and SC-0 examples do not share similar properties. This suggests that knowledge of the differences between typical subjects and typical objects is insufficient to resolve the DO/SC ambiguity. Instead, if comprehenders are to exploit the sources of information revealed in Study 1, they must have more specific knowledge about the properties that distinguish NPs that form typical objects of the verbs in question and the NPs that form typical embedded subjects of complementizer-less clauses for these verbs. Additionally, these data suggest that not only are the embedded subjects in the *SC-that* and SC-0 examples different, but that the subjects of the *SC-that* examples are in fact similar to the direct objects of the DO examples with respect to the properties measured by our model. We also find that although different verbs seem to have a preference for either SC-0 or *SC-that* sentential complements, the differences in verb distribution

between these two complement types at most only partially account for the failure of the DO/SC-0 trained model to correctly identify the subcategorization of the pseudo-ambiguous SC-*that* examples.

### Study 3: Predicting SC-0/SC-*that* Subcategorization

The results from Study 2 pose a potential problem for our interpretation of the success of the first regression model as indicating that there is a large quantity of information available for resolving the DO/SC ambiguity. Specifically, our second set of results suggests that the pseudo-ambiguous embedded subjects bear a strong resemblance to the post-verbal NPs of the direct object examples, whereas the embedded subjects of the examples without the complementizer *that* do not share this resemblance. This raises the possibility that the success of our model in resolving the DO/SC ambiguity is due to an effort on the part of the speakers/writers of our corpus data to avoid ambiguity by using the complementizer *that* whenever there is the potential for ambiguity, while leaving out the complementizer *that* only in those cases where there is sufficient information available to resolve the ambiguity. This would suggest that the high rate of resolving the DO/SC ambiguity was not due to the ambiguity resolution ability of the model, but to the ambiguity avoidance skills of the producers of the input data.

In order to investigate this possibility, we perform a third set of analyses by using our corpus data and regression models to examine the factors that govern the use of the complementizer *that* in sentential complements. We used a regression model to directly predict the presence or absence of the complementizer *that* in sentential complement uses of the 100 verbs listed in Garnsey et al. (1997). This makes it possible to measure both the degree to which complementizer usage is predictable in natural language use (based on a much larger sample than previous studies), and to compare the relative strengths of the various predictors of complementizer use.

#### *Method*

This study used the binary logistic regression model used in the previous studies, as well as the coded corpus data from those studies. Specifically, we used the 68,016 SC-0 and 74,940 SC-*that* BNC corpus examples of the 100 Garnsey et al. (1997) verbs, coded for the length, semantic, frequency, and lemma factors described above. We excluded all examples with an unambiguously case-marked pronoun as head of the post-verbal NP. All information was entered into the binary logistic regression model in a single step.

#### *Results*

We report two sets of results. The first set of results reflects the overall performance of the model when all information is entered in to the regression in a single step. These results reflect the combined contribution of all sources of information. The second set of results reflects the performance of the model when only specific subsets of the data are used. These results reflect the amount of information contributed by these separate sources of information.

*Overall Contribution of all Sources of Information*

The results of the regression model for predicting SC-0 vs. *SC-that* are shown in Table 4. The baseline of always predicting *SC-that* (the most frequent case) is 52.4%. When all length, frequency, semantic, and lemma predictors are included, the final level of performance is 78.2%, indicating that the presence or absence of the complementizer *that* can be predicted in natural language use more than three quarters of the time.

We evaluated the performance of the regression model when given only certain subsets of input information in order to evaluate (a) how much information is available at different points during sentence processing, and (b) how much information is available from different types of information (frequency, length, semantic, lemma). As with the complete regression model, the baseline performance is 52.4%. The contributions of each source of information are shown in Table 4.

Predictors included	Baseline %	Model %	Nagelkerke R square	Model Chi Square
Main model (all info)	52.4%	78.2%	0.487	$\chi^2$ (186, N = 142956) = 64892, p<0.001
Lemma only	52.4%	75.4%	0.408	$\chi^2$ (98, N = 142956) = 52184, p<0.001
Semantic only	52.4%	74.2%	0.367	$\chi^2$ (80, N = 142956) = 46021, p<0.001
Length only	52.4%	69.4%	0.24	$\chi^2$ (5, N = 142956) = 28309, p<0.001
Frequency only	52.4%	68.2%	0.206	$\chi^2$ (3, N = 142956) = 24023, p<0.001
Subject info only	52.4%	72.2%	0.282	$\chi^2$ (43, N = 142956) = 33960, p<0.001
Post-verbal NP info only	52.4%	69.6%	0.259	$\chi^2$ (45, N = 142956) = 30805, p<0.001
Post-verbal NP length only	52.4%	65.4%	0.087	$\chi^2$ (1, N = 142956) = 9666, p<0.001

Table 4. Binary Logistic Regression Models for predicting SC-0 vs. *SC-that*.

*Contributions of Various Subsets of Information*

The contribution of this information can again be organized in two different ways, as in Study 1. First, contributions can be organized in terms of the type of information, namely, lemma, semantic, length, and frequency information (see Figure 5). This analysis reveals that like in the DO/SC analyses, verb lemma identity was the richest source of information, adding an additional 23.0% correct to the performance of the model. However, even the weakest additional source of information, frequency information, still added 15.8% correct to the performance of the model. All specific types of information contributed significant additional variance to the model's predictions.

Alternatively, contributions can be organized in terms of where the information arises in the sentence when processing from beginning to end, by considering the contributions of the factors associated with the main subject, the lemma, and the post-verbal NP respectively. These data are shown in Figure 6. Here, unlike the DO/SC models, the richest source of information (apart from lemma information) was the main-subject NP, adding 19.8% correct to the performance of the model (recall that in the DO/SC models, main-subject NP information was the weakest source). Post-verbal NP information added 17.2% correct to the performance of the model, still a significant contribution.

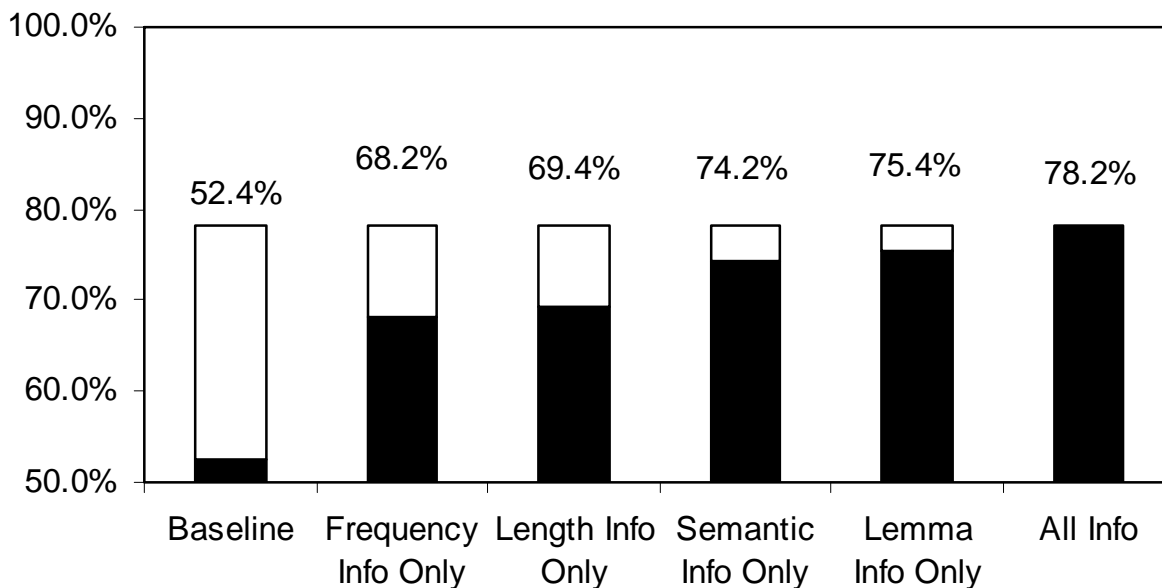


Figure 5. Performance of SC-0/SC-that regression model with information type based subsets of information.

### Discussion

The results in Study 3 suggest that there is abundant information available in the context to predict whether a sentential complement is likely to have an overt complementizer or not. As in the case of the SC-0/DO model, the identity of the verb (lemma) is the most important of the predictors. This corresponds with the observations of several norming studies (e.g., Garnsey et al., 1997; Trueswell et al., 1993) that different verbs have different *that* preferences. In other words, because some verbs are much more likely than others to take a complement with the *that*, knowing the target verb improves the ability to predict the presence or absence of *that*. In fact, even though the overall bias in our input data is (roughly) equally split between SC-*that* and SC-0, three quarters of the individual verbs actually have an SC-*that* bias. The overall SC-0 bias arises primarily from a small number of high-frequency SC-0 biased verbs. The verb *think* alone accounts for 40.6% of our SC-0 observations, while the verbs *think*, *know*, *believe*, *hope*, *feel*, and *suppose* account for 73.5% of the SC-0 observations. These numbers are similar to those obtained from the larger set of BNC data when all SC-*that* and SC-0 examples are taken into account, not just those with ambiguous pronouns.

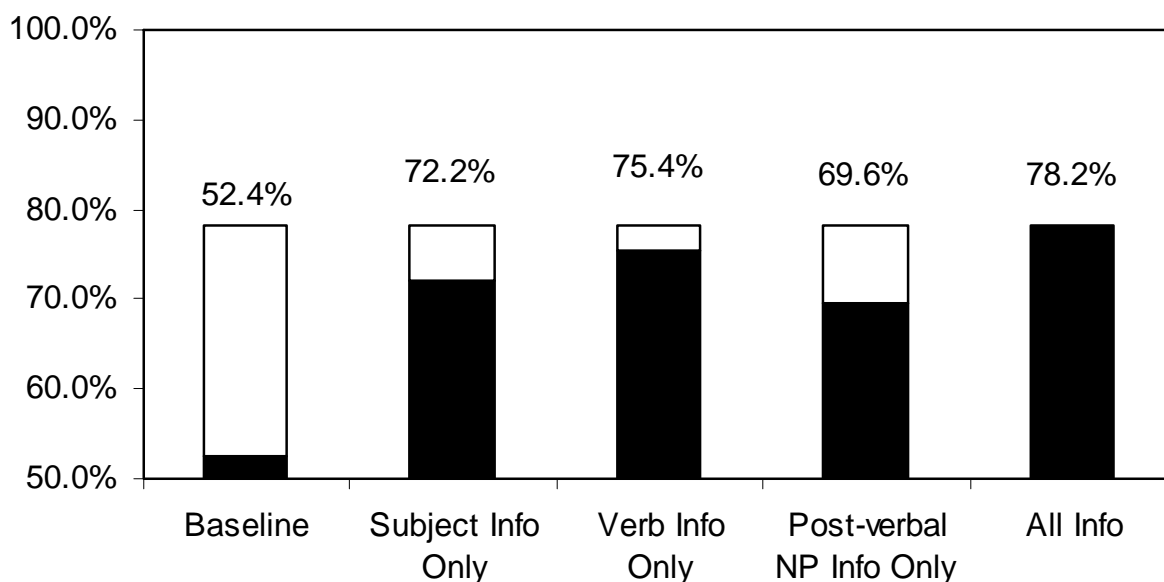


Figure 6. Performance of SC-0/SC-that regression model with location based subsets of information.

The pattern of verbs which prefer to appear without the complementizer fits in with the observations in Thompson and Mulac (1991). They provide evidence that “first and second person subjects, the verbs *think* and *guess*, pronominal complement subjects, and auxiliaries, indirect objects, and adverbs are significant in predicting the use of *that*.” They argue that all of these factors are related, because they all correspond with a weakening of the distinction between main and complement clause, either because they correspond with the subject of the subordinate clause being the discourse topic or the reduction of the main subject and verb to an epistemic use (the difference between *I believe* in *I believe in truth and justice* and *I believe it’s going to rain*). The verbs listed above, *think*, *know*, *believe*, *hope*, *feel*, and *suppose*, are all verbs that likely to occur in epistemic usages.

The second most important set of predictors for determining the presence of the complementizer *that* were the semantic predictors. With just semantic information, the model is able to achieve a 74.2% performance level, as shown in Figure 5. As in Study 1, we inspected a ranked list of subjects and post-verbal NPs in order to assign post hoc interpretations to the type of word or NP that the model preferred for each continuation. The general tendency was for examples that were most strongly predicted to be SC-0 to have animate and pronominal post-verbal NPs, while the examples that were most strongly predicted to be SC-*that* tended to have post-verbal NPs that were more like the DO post-verbal NPs – full NPs, less animate, and more abstract.

Length information also plays a role in predicting complementizer use. Post-verbal NP length has previously been argued to affect the presence of the complementizer (e.g., Hawkins, 2002). Hawkins argues for a principle of *Maximize On-Line Processing*, which predicts “a preference for [the use of the complementizer *that*], in direct proportion to the length of the subordinate subject”. He analyzes corpus data from Rohdenburg (1999) to show that when the

post-verbal NP is a personal pronoun, the use of the complementizer *that* is at its lowest, while it increases with one and two word full NPs, and is highest when the post-verbal NP has three or more words.

Target	SC_0		SC_That		Two-tailed significance
	Mean	SD	Mean	SD	
Subject Head Noun	2.6	2.3	4.3	3.0	t(142954) = 125.6, p < .001
Post-verbal NP first word	3.4	1.8	4.1	2.3	t(142954) = 64.8, p < .001
Post-verbal NP head noun	4.2	2.4	6.0	2.7	t(142954) = 131.5, p < .001
Subject NP	4.8	10.7	11.0	17.9	t(142954) = 78.6, p < .001
Post-verbal NP	9.4	15.6	18.1	20.9	t(142954) = 89.1, p < .001

Table 5. Mean lengths in characters of various elements of *SC-that* and SC-0 examples.

We found that, although length was less important than either verb identity or semantic factors, it still played an important role in predicting the presence of *that*. As Figure 5 shows, all length factors combined produced a performance level of 69.4%. All five length measures (subject NP length, subject NP head length, post-verbal NP length, post-verbal NP head length, and post-verbal NP first word length) showed a significant difference between SC-0 and *SC-that* (see Table 5). In all cases except for the length of the subject NP, the longer length was associated with the presence of a complementizer. Because of the proposed role of the length of the post-verbal NP in theories such as that proposed by Hawkins (2002), we also tested the performance of the regression model when given the length of the post-verbal NP alone. This produced a performance level of 65.4%, which, though significant (Model Chi Square (1, N = 142956) = 9666.616, p < 0.001), accounts for only about 9% of the variance (Nagelkerke R square = .087). Thus, length is clearly part of the story, but not the whole story.

	Full post-verbal NP	Pronominal post-verbal NP	% Pronominal post-verbal NP
DO	193206	50877	20.8%
SC-That	91691	32565	26.2%
SC-0	49543	87228	63.8%

Table 6. Frequencies of post-verbal NP pronominalization by subcategorization.

We are also able to break down the contributions of different sources of information by the part of the sentence that they come from. The contribution of the verb lemma is discussed above, leaving for consideration the information contained in the post-verbal NP and the information contained in the main subject. As in the case of the DO/SC-0 model, we found that the post-verbal NP also contains a significant amount of information for predicting, in this case, the presence or absence of the complementizer *that*. One potential source of information from the post-verbal NP for both the DO/SC-0 and the SC-that/SC-0 prediction task is the pronominal status of the post-verbal NP. As Table 6 shows<sup>11</sup>, the post-verbal NPs for SC-0 examples are more likely to be a pronouns, while the post-verbal NPs for both SC-that and DO examples are

<sup>11</sup> For purposes of general informativeness, the data in Table 6 includes examples where the post-verbal NP contains an unambiguously case marked pronoun, which were excluded from the model input. Additionally, the data in Table 6, Table 7, and Table 8 differ from the model input in that the model input excludes all examples for which we could not generate a complete set of predictor values (typically due to a failure to generate a semantic vector for one of the elements or a failure on the part of the algorithm used to identify the head noun of a constituent).

more likely to be full NPs. Although the full NP vs. pronoun status of the post-verbal NP is not coded for directly in the regression model, but is reflected in measures such as length, frequency, and potentially semantics.

One might take the data presented in Table 6 to suggest that the complementizer *that* was specifically used as a disambiguation device when the post-verbal NP was a full NP, and thus more likely to be confused with a direct object. However, there is evidence to suggest that this is not likely to be the case. Although the regression model input data only contains cases where any post-verbal NP pronoun is not unambiguously case marked, the corpus data which we used contains many examples where the post-verbal NP consists of an unambiguously case marked pronoun. If the complementizer *that* were used as a disambiguation device, one would expect that the ambiguous pronouns would be more likely to be disambiguated with *that* than the unambiguous pronouns. As shown in Table 7, this is not the case. In fact, the complementizer is even more likely to be omitted when the pronoun is ambiguous. This is consistent with experimental results from Ferreira and Dell (2000). They provide evidence that in controlled, minimally varying sentences (the kind that are typically used in psycholinguistic experiments), comprehenders do not use the complementizer to avoid ambiguity (and may not even be aware of the ambiguity as they produce sentences).

Unlike the case of the SC/DO ambiguity, where we found that the post-verbal NP played a larger role than the subject NP, we found that role played by the subject NP was slightly larger, with the subject-information-only model, shown in Figure 6, having a 72.2% performance level, and the post-verbal NP information only model having a 69.9% performance level.

Type of post-verbal NP	SC-0	SC-That	% <i>that</i> ellipsis
Full NP	49543	91691	35.1%
Unambiguous pronoun	49754	19855	71.5%
Ambiguous pronoun	37474	12710	74.7%

Table 7. Frequencies of *that* ellipsis by type of post-verbal NP.

The relative importance of the main clause subject in predicting the presence or absence of the complementizer *that* seems, at least in part, to be related to the epistemic nature of the *that*-less sentential complements described in Thompson and Mulac (1991). They provide data showing that the complementizer *that* is much more likely to be absent when the main clause subjects are either *I* or *you* than it is for other main clause subjects. We replicate this finding in a separate analysis of our corpus data. Table 8 shows the BNC frequencies of SC-*that* and SC-0 for the 100 Garnsey et al. verbs separated by main clause subject. Unlike the regression model data used in this paper, the data in Table 8 does not exclude examples where the post-verbal NP contains an unambiguously case marked pronoun.

Main Subject	SC-0	SC-That	% <i>that</i> ellipsis
I	14832	3271	82%
You	12818	2620	83%
All Others	109121	118365	48%

Table 8. Frequencies of *that* ellipsis by main subject.

Finally, Ferreira and Dell (2000) present evidence for contextual factors influencing the use of the complementizer *that* in sentence production. Their evidence suggests that *that* is omitted more often when the subsequent material was either repeated from earlier in the sentence or was prompted by a recall cue, and thus the reduction in the use of *that* results from an effort to allow the early mention of available material. We find that there are many cases in the corpus data where the main clause subject is repeated as the embedded subject. These cases are primarily when the *I* or *you* main clause subject is repeated as the embedded subject, and thus it is difficult to distinguish between epistemicity as a cause for the lack of complementizer, and Ferreira and Dell's early mention.

### *Conclusions for Study 3*

We find that the use of the complementizer *that* is governed by highly predictable factors that appear to be unrelated to the avoidance of ambiguity per se. In fact, the producers of the corpus data actually use the complementizer *less* frequently in some ambiguous contexts than they do in similar unambiguous contexts. This fits in with evidence from Ferreira and Dell (2000) that suggests that speakers may not even be aware of potential ambiguities during production.

Instead, it appears that the factors determining complementizer use are related to other factors, such as epistemicity, as described in Thompson and Mulac (1991). This suggests that although the problem of resolving the DO/SC ambiguity is not made easier by means of cooperative producers of language, it may be made easier than it otherwise would by a conspiracy of other factors governing the contexts in which the complementizer is most commonly omitted, and therefore, the contexts in which the DO/SC structural ambiguity most frequently occurs. However, it may be possible that diachronic language-evolution pressures only allow structural ambiguities to arise in circumstances where there is usually enough non-structural information available for resolving the ambiguity.

### Conclusions

Our combined results suggest that not only is there abundant information available for resolving the DO/SC ambiguity, but that there is also a large amount of information available for predicting whether the complementizer *that* will be used in a given instance. This information includes both information about the semantic nature of the post-verbal NP (related to phenomena such as thematic fit, e.g., McRae et al., 1998), and information about the structural properties (full NP vs. pronoun, length, frequency). Our results also suggest that although there is some overlap in these sources of information, each source does have the potential to make a unique contribution to the final resolution of the ambiguity. Additionally, we assume that the types of information captured by our model underrepresent the types of information actually available, due to the relative simplicity of the model.

The large amount of contextual information available for resolving the DO/SC ambiguity and for prediction complementizer presence/absence in sentential complements suggests that comprehenders could in principle not only generate expectations about whether a particular context is leading up to a direct object or a sentential complement, but also about whether or not that sentential complement will have a complementizer. In fact, there is some anecdotal evidence

from DO/SC reading time experiments that the post-verbal NP takes longer to process in complementizer-less contexts than it does in equivalent examples with complementizers, but only when subjects are biased towards expecting a sentential complement (and, presumably, the complementizer *that*). However, these experiments were not explicitly designed to look for evidence of *that* expectations.

Our results also suggest that a complex set of factors need to be taken into account during processing. Study 2 demonstrates that knowledge of the types of noun phrases that typically occur as subjects and objects is insufficient for resolving the DO/SC ambiguity. Instead, the comprehender must rely on a more complex notion of how likely a particular NP is to be a subject or an object in a given context. However, our results also suggest that there is a rich source of information available for resolving the DO/SC ambiguity. In turn, this portrays a linguistic signal that is rich with information for guiding structural analysis. Linguistic expressions are more than just the words that constitute them; rather, they unfold in patterns that reflect the meanings and structures that speakers are trying to convey.

## References

- Altmann, G. T. M. (1998). Ambiguity in sentence processing. *Trends in Cognitive Sciences*, 2, 146-152.
- Altmann, G. T. M. (1999). Thematic role assignment in context. *Journal of Memory & Language*, 41(1), 124-145.
- Charniak, E. (1997). *Statistical parsing with a context-free grammar and word statistics*. AAAI-97, Providence, RI.
- Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by Latent Semantic Analysis. *Journal of the American Society For Information Science*, 41, 391-407.
- Ferreira, F., & Henderson, J. M. (1991). Recovery from misanalyses of garden-path sentences. *Journal of Memory & Language*, 30(6), 725-745.
- Ferreira, V. S., & Dell, G. S. (2000). Effect of ambiguity and lexical availability on syntactic and lexical production. *Cognitive Psychology*, 40(4), 296-340.
- Frazier, L. (1978). *On comprehending sentences: syntactic parsing strategies*. Unpublished PhD, University of Connecticut.
- Garnsey, S. M., Pearlmutter, N. J., Myers, E., & Lotocky, M. A. (1997). The contributions of verb bias and plausibility to the comprehension of temporarily ambiguous sentences. *Journal of Memory & Language*, 37(1), 58-93.
- Hare, M., McRae, K., & Elman, J. (2003). Sense and structure: Meaning as a determinant of verb subcategorization preferences. *Journal of Memory and Language*, 48, 281-303.
- Hawkins, J. A. (2002). Symmetries and asymmetries: their grammar, typology and parsing. *Theoretical Linguistics*, 28(2), 95-150.
- Juliano, C., & Tanenhaus, M. K. (1993). Contingent frequency effects in syntactic ambiguity resolution, *Proceedings of the 15th Annual Conference of the Cognitive Science Society* (pp. 593-598). Mahwah, NJ: Lawrence Erlbaum Associates.
- Kintsch, W. (2001). Predication. *Cognitive Science*, 25, 173-202.
- Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The Latent Semantic Analysis theory of the acquisition, induction, and representation of knowledge. *Psychological Review*, 104, 211-240.
- Landauer, T. K., Foltz, P. W., & Laham, D. (1998). Introduction to Latent Semantic Analysis. *Discourse Processes*, 25, 259-284.
- Landauer, T. K., Laham, D., & Foltz, P. W. (1998). Learning human-like knowledge by Singular Value Decomposition: A progress report. In M. I. Jordan & M. J. Kearns & S. A. Solla (Eds.), *Advances in Neural Information Processing Systems 10* (pp. 45-51). Cambridge: MIT Press.
- Landauer, T. K., Laham, D., Rehder, B., & Schreiner, M. E. (1997). *How well can passage meaning be derived without using word order? A comparison of Latent Semantic Analysis and humans*. Proceedings of the 19th annual meeting of the Cognitive Science Society.
- MacDonald, M. C., Pearlmutter, N. J., & Seidenberg, M. S. (1994). The lexical nature of syntactic ambiguity resolution. *Psychological Review*, 101(4), 676-703.
- MacWhinney, B., & Bates, E. (1989). *The Crosslinguistic Study of Sentence Processing*. Cambridge; New York: Cambridge University Press.

- McRae, K., Spivey-Knowlton, M. J., & Tanenhaus, M. K. (1998). Modeling the influence of thematic fit (and other constraints) in on-line sentence comprehension. *Journal of Memory & Language*, 38(3), 283-312.
- Rehder, B., Schreiner, M. E., Wolfe, M. B., Laham, D., Landauer, T. K., & Kintsch, W. (1998). Using Latent Semantic Analysis to assess knowledge: Some technical considerations. *Discourse Processes*, 25, 337-354.
- Rohdenburg, G. (1999). Clausal Complementation and Cognitive Complexity in English. In F.-W. Neumann & S. Schylting (Eds.), *Anglistentag 1998 Erfurt: Proceedings* (pp. 101-112). Trier, Germany: Wissenschaftlicher.
- Roland, D., Dick, F., & Elman, J. L. (2003). Frequency of basic English grammatical structures: A corpus analysis. Manuscript submitted for publication.
- Spivey, M. J., & Tanenhaus, M. K. (1998). Syntactic ambiguity resolution in discourse: Modeling the effects of referential context and lexical frequency. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, 24(6), 1521-1543.
- Thompson, S. A., & Mulac, A. (1991). The discourse conditions for the use of the complementizer 'that' in conversational English. *Journal of Pragmatics*, 15, 237-251.
- Trueswell, J. C., Tanenhaus, M. K., & Kello, C. (1993). Verb-specific constraints in sentence processing: Separating effects of lexical preference from garden-paths. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, 19(3), 528-553.
- Wolfe, M. B., Schreiner, M. E., Rehder, B., Laham, D., Foltz, P. W., Kintsch, W., & Landauer, T. K. (1998). Learning from text: Matching readers and text by Latent Semantic Analysis. *Discourse Processes*, 25, 309-336.

Author Note

Douglas W. Roland, Jeffrey L. Elman, and Victor S. Ferreira, Center for Research in Language, University of California, San Diego

This research was funded by NIH/NIDCD5T32DC00041, NIH/NIMHR01-MH60517, and NIH/NIMHR01-MH64733.

Correspondence should be addressed to Douglas Roland, Center for Research in Language, University of California, San Diego, 9500 Gilman Drive, Dept. 0526, La Jolla, California 92093-0526. Email: [droland@crl.ucsd.edu](mailto:droland@crl.ucsd.edu)