Expectation-based syntactic comprehension

Roger Levy
Department of Linguistics
University of California, San Diego

May 20, 2007

Abstract

This paper investigates the role of resource allocation as a source of processing difficulty in human sentence comprehension. The paper proposes a simple information-theoretic characterization of processing difficulty as the work incurred by resource reallocation during parallel, incremental, probabilistic disambiguation in sentence comprehension, and demonstrates its equivalence to the theory of Hale (2001), in which the difficulty of a word is proportional to its *surprisal* (its negative log-probability) in the context within which it appears. This proposal subsumes and clarifies findings that high-constraint contexts can facilitate lexical processing, and connects these findings to well-known models of parallel constraint-based comprehension. In addition, the theory leads to a number of specific predictions about the role of expectation in syntactic comprehension, including the reversal of locality-based difficulty patterns in syntactically constrained contexts, and conditions under which increased ambiguity facilitates processing. The paper examines a range of established results bearing on these predictions, and shows that they are largely consistent with the surprisal theory.

Keywords:
Parsing; Frequency; Sentence Processing; Information Theory;
Prediction; Syntax; Word Order; Syntactic Complexity

1 Introduction

There are several important properties that must be accounted for by any realistic theory of human sentence comprehension. These include:

1. robustness to imperfectly formed input;
2. accurate ambiguity resolution;
3. inference on the basis of incomplete input; and

4. differential, localized processing difficulty.

This paper attempts to show how these four properties can be tightly interconnected in a probabilistic, expectation-based theory of syntactic comprehension. In particular, this paper focuses on deriving a theory of Property 4—namely, that not all sentences are equally easy to comprehend, and different parts of sentences differ in their difficulty—from Properties 1 through 3.

To a considerable extent, the dominant paradigm for investigating differential processing difficulty has been what I will call *resource-requirement* or *resource-limitation* theories. These propose that:

- some syntactic structures require more of a given resource than do others; and
- that resource is in short supply in the human parser; and
- this gives rise to greater processing difficulty for more resource-intensive structures.

Typically this limited resource is some form of *memory*. The resource-limitation position has also come to inform a persistent view of ambiguity resolution: the resource-limited parser can only pursue one alternative at a time (i.e., the parser is serial), and in the face of local ambiguity, the processor chooses the alternative that minimizes the resources consumed. This viewpoint has inspired a variety of ambiguity resolution theories, including Late Closure (Frazier and Fodor, 1978) and Minimal Attachment (Frazier, 1979). Perhaps the most salient modern incarnations of memory-centered resource-requirement theories are, for ambiguity resolution, the Active Filler Hypothesis (AFH; Clifton and Frazier 1989); and, for locally unambiguous sentences, the Dependency Locality Theory (DLT; Gibson 1998, 2000).

At the same time, an alternative line of research has focused on the role of expectations in syntactic processing. This idea has historically been associated most closely with *constraint-satisfaction* processing models such as those of MacDonald (1993); MacDonald et al. (1994); Tanenhaus et al. (1995), and McRae et al. (1998), and can be traced back to early work by Marslen-Wilson (1975).[1] This line of work typically takes a strong integrationist and parallelist perspective: the comprehender draws on a variety of information sources (structural, lexical, pragmatic, discourse) to evaluate in parallel a number of possible alternatives for the input seen thus far. For the most part, the primary concern of constraint-based work has been ambiguity resolution, the argument being that possible structural analyses are ranked according to their plausibility on a number of dimensions, rather than according to the amount of resources they consume. Empirically observed processing difficulty after local ambiguity resolution is informally ascribed to either a *reranking* of the favored analysis, or *competition* between closely-ranked analyses. The constraint-based position can be thought of as a *resource-allocation* approach to syntactic processing: the parser allocates different amounts of resources to different interpretations of the partial input, and difficulty arises when those resources turn out to be inefficiently allocated.

[1] See Jurafsky (2003) for a more comprehensive account of the history of expectation-based approaches in human sentence processing, including syntactic processing.

As argued by Jurafsky (2003), probability theory fits naturally as an underlying infrastructure for constraint-based approaches to express the rational (in the sense of Anderson 1990) combination of multiple information sources. The use of probability theory for psycholinguistic modeling has in fact become more prevelant over the past decade, beginning with Jurafsky (1996) and continuing in Narayanan and Jurafsky (1998, 2002); Crocker and Brants (2000). This paper proposes a resource-allocation theory of processing difficulty grounded in parallel probabilistic ambiguity resolution: the possible structural analyses consistent with a partial input are preferentially ranked in parallel, and the difficulty of a new word corresponds to the amount of reallocation necessary to reflect the word's effect on the preference ranking. Section 2 gives the derivation of this theory and shows that it turns out to be equivalent to the *surprisal* theory originally proposed by Hale (2001).[2] As a result we have a single theory (simply called the *surprisal theory* in this paper) unifying the idea of the work done incremental probabilistic disambiguation with expectations about upcoming events in a sentence. In this theory, surprisal serves as a causal bottleneck between the linguistic representations constructed during sentence comprehension and the processing difficulty incurred at a given word within a sentence. This paper argues that the surprisal theory, when conjoined with probabilistic models chosen according to appropriate principles (see Section 3), makes a wide range of precise predictions consistent with empirical observations, while remaining relatively neutral as to the exact representations of possible structural analyses. Section 4 contrasts the surprisal theory with alternative resource-allocation and resource-limitation theories of processing difficulty, illustrating the general conditions under which their predictions maximally diverge. The remainder of the paper examines a number of established experimental results pertaining to these divergent predictions, and shows that they lend considerable support to the surprisal theory.

2 Deriving a resource-allocation theory of processing difficulty

This section presents a new derivation of a theory of resource-allocation processing difficulty, based on a highly general conception of sentence comprehension, and accounting for principles that are necessary for any realistic model of human sentence processing.

A language contains a (normally infinite) set of *complete structures* such that a fully disambiguated utterance corresponds to exactly one structure. Each structure contains the complete string of the utterance, plus presumably at least some other information, since some well-formed strings are ambiguous. As an example, we might consider a complete structure to be the string plus its syntactic/semantic analysis, so that the sentence *the girl saw the boy with a telescope* might be compatible with two possible complete structures, one

[2]The surprisal theory of Hale (2001) is not to be confused with the Entropy Reduction Hypothesis (ERH) of Hale (2003b,a, 2006). In the former, the difficulty of a word is determined by its log-probability; in the latter, by the induced change in uncertainty as to the complete analysis of the sentence. These two quantities need not be related.

where *with a telescope* modifies *saw* and one where it modifies *boy*. However, we will remain agnostic as to precisely what these complete structures contain, so long as they contain the complete string.

We can reasonably define what it means to *comprehend* a sentence S as the (implicit or explicit) construction of a preference ranking over the set of possible all possible structures $\mathcal{T}$ in the language consistent with S. We will use the language of probability theory to express preferences and rankings, so comprehension of S involves placing a probability distribution over $\mathcal{T}$ once we have seen S.

There is ample evidence, however, that sentence comprehension is *incremental*: we do not wait until we have heard an entire sentence to start disambiguating and comprehending. Perhaps the most explicit demonstration of this fact comes from work in cross-modal eye-tracking (Tanenhaus et al., 1995; Altmann and Kamide, 1999; Kaiser and Trueswell, 2004); in Altmann and Kamide (1999), for example, listeners were found to start looking at the plausible objects in a picture for the main verb of a sentence as soon as they heard the verb. Comprehenders are able to make inferences about later parts of the sentence based on what they have heard earlier in the sentence. To capture this fact, we define the comprehension of a *partial* input sequence $w_{1\ldots i}$ (the first i words of the sentence) to be placing a preference (i.e., probability) distribution P_i^T over the possible structures T based on $w_{1\ldots i}$, plus context external to the sentence itself. For listeners to be capable of incremental inference, they must be constantly updating P_i^T; for simplicity in the present context, we assume that they update P_i^T after every input word.

The probability distribution P_i^T consists of an allocation of resources among the possible interpretations of the sentence, and for the resource-allocation theory of processing difficulty our single stipulation will be that difficulty is incurred by updating P_i^T, and that difficulty is quantified by the degree that P_i^T has to be updated. To quantify the degree of difficulty in the update we will use the *relative entropy* of the updated distribution with respect to the old distribution.[3] The relative entropy of a probability distribution q with respect to another distribution p (also known as the *Kullback-Leibler (KL) divergence* of q from p) is defined as

$$D(q||p) = \sum_{T \in \mathcal{T}} q(T) \log \frac{q(T)}{p(T)} \tag{1}$$

Intuitively speaking, the relative entropy of q with respect to p can be thought of as the penalty incurred from encoding the distribution q with p. When $q = p$, $D(q||p) = 0$, and the greater the difference between the distributions, the greater the relative entropy.

It turns out that under this formulation of resource-allocation processing difficulty, regardless of the form of complete structures T or the preference distribution P^T, the predicted difficulty of the i^{th} word, w_i, is precisely equal to the *surprisal* of w_i, which is defined as the negative log-probability of w_i in its sentential context (which we denote by the

[3]It is of interest to note that recently, researchers in vision have independently proposed the relative entropy induced by an observation as a theoretical quantification of what drives attention in human visual scene perception (Itti and Baldi, 2005).

already-seen input sequence $w_{1\cdots i-1}$) and extra-sentential context (which we denote simply by CONTEXT):

$$\textit{difficulty} \;\propto\; -\log P(w_i|w_{1\cdots i-1}, \text{CONTEXT}) \tag{2}$$

Precisely this measure of difficulty was in fact proposed by Hale (2001). Surprisal is minimized (goes to zero) when a word *must* appear in a given context (i.e., when $P(w_i|w_{1\ \cdots\ i-1}, \text{CONTEXT}) = 1$), and approaches infinity as a word becomes less and less likely. The simple proof of this result is given in Section 2.1, and its implications are discussed in Section 2.2.

2.1 Proof of equivalence to surprisal

Consider any stochastic generative process P, conditioned on some (possibly null) external context, that generates complete structures $T \in \mathcal{T}$, each consisting at least partly of surface strings to be identified with serial linguistic input. Examples of such processes include but are not limited to n-gram models, Hidden Markov Models (HMMs), and probabilistic context-free grammars (PCFGs). Furthermore, for any particular input prefix $w_{1\cdots i}$ define the probability distribution P_i as the conditional distribution over $\mathcal{T}$ induced by P, given the prefix $w_{1\cdots i}$ and other context:[4]

$$P_i(T) \equiv P(T|w_{1\cdots i}), \forall T \in \mathcal{T} \tag{3}$$

and define the set $\mathcal{T}_i$ as the set of complete structures with prefix $w_{1\cdots i}$ (note that $\mathcal{T}_i$ is also the subset of $\mathcal{T}$ that has non-zero probability according to P_i). We will also give P and P_i a secondary meaning as signifying joint and conditional (respectively) probability distributions over words: $P(w_{1\cdots i}) \equiv \sum_{T\in\mathcal{T}_i} P(T)$, and $P_i(w) \equiv P(w|w_{1\cdots i})$.

I will now show that

$$D(P_{k+1}||P_k) = -\log P_k(w_{k+1}) \tag{4}$$

That is, the relative entropy of the distribution over hidden structures *after* having seen w_{k+1} from the distribution *before* having seen w_{k+1} is simply equal to the surprisal of w_{k+1}.

Proof. The proof requires only a simple application of the chain rule. First, note that for any integer j and any $T \in \mathcal{T}_j$,

$$P_j(T) \equiv P(T|w_{1\cdots j}) \tag{5}$$

$$= \frac{P(T, w_{1\cdots j})}{P(w_{1\cdots j})} \tag{6}$$

[4]For convenience we will omit the CONTEXT term explicitly conditioned on in Equation (2), but it should be understood that we are always implicitly conditioning on extra-sentential context.

And by virtue of the fact that T is in $\mathcal{T}_j$,

$$P_j(T) = \frac{P(T, w_{1\cdots j})}{P(w_{1\cdots j})} \quad (7)$$

$$= \frac{P(T)}{P(w_{1\cdots j})} \quad (8)$$

Therefore, for all $T \in \mathcal{T}_{k+1}$,

$$\frac{P_{k+1}(T)}{P_k(T)} = \frac{\frac{P(T)}{P(w_{1\cdots k+1})}}{\frac{P(T)}{P(w_{1\cdots k})}} \quad (9)$$

$$= \frac{P(w_{1\cdots k})}{P(w_{1\cdots k+1})} \quad (10)$$

$$\equiv \frac{1}{P_k(w_{k+1})} \quad (11)$$

independent of T.

Therefore, the KL divergence from P_{k+1} to P_k is

$$D(P_{k+1}||P_k) = \sum_{T \in \mathcal{T}_{k+1}} P_{k+1}(T) \log \frac{P_{k+1}(T)}{P_k(T)} \quad (12)$$

$$= \log \frac{1}{P_k(w_{k+1})} \sum_{T \in \mathcal{T}_{k+1}} P_{k+1}(T) \quad (13)$$

$$= -\log P_k(w_{k+1}) \quad (14)$$

□

Intuitively, this proof results from the fact that the ratio of the probability of any complete structure T before versus after seeing a word w_{k+1} is constant, because the original process generating $\mathcal{T}$ is the same. This constant ratio has to be the amount of probability mass pruned away from P_k by the requirement of compatibility with w_{k+1}—in other words, the conditional probability of w_{k+1}, as seen in Equation 11. This is the probability ratio term in the KL divergence, as seen in Equation 13, and because it is constant, the probability over structures T can be independently summed out. Finally, note that this proof of equivalence only holds if the extrasentential context does not change at the same time as w_{k+1} is processed; if the extrasentential context is changed, the constancy of the ratio $\frac{P_{k+1}(T)}{P_k(T)}$ in 11 may be broken.

2.2 Implications of relative-entropy derivation of surprisal

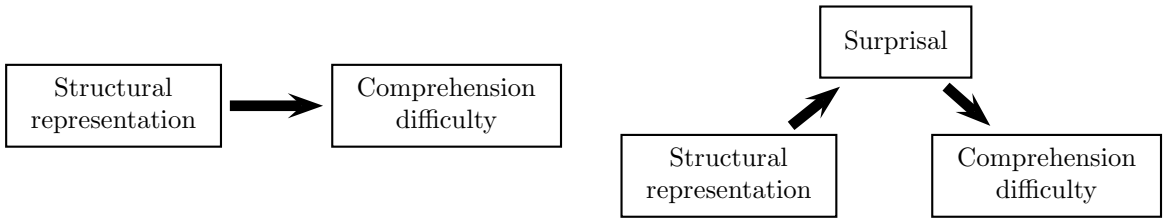
This equivalence has important implications for how we conceptualize the incremental parsing process. In a fully parallel, incremental probabilistic parser capable of online inference (that is, inference before input is complete), storing the complete set of ranked partial parses consistent with already-seen input is also equivalent to assigning a probability distribution over the complete structures to which the already-seen input may possibly extend. Upon termination of the input, this set of ranked partial parses determines a most-likely interpretation. On the way, after every new input token, such a parser must update its collection of ranked partial parses—and therefore its distribution over completed parses—to reflect the new information. Intuitively, the relative entropy from a distribution p to another distribution q measures the penalty incurred by encoding, or approximating, q with p. The surprisal can therefore be interpreted as the difficulty incurred in replacing the old distribution with the new. A word's surprisal is also, of course, a measure of its expectancy.

Deriving surprisal as a special case of reranking-based difficulty thus addresses a potential conceptual vulnerability of expectation-based approaches. It might be thought that calculating expectations about upcoming structures in a sentence can be computationally expensive, so why would the human parser waste resources on constantly calculating and updating the likelihood of upcoming words and/or structures in a sentence? For those who are inclined to think of incremental structure-building and disambiguation as the fundamental type of work that needs to be done in sentence comprehension, we now have a clear answer to this challenge: surprisal as the predicted difficulty of word w_i falls out of the incremental update process itself. Expectations about upcoming words in a sentence need not be explicitly calculated; rather, they are implicit in the partial parse of an incomplete input.

2.3 Surprisal as a causal bottleneck

The hypothesis that the surprisal of a word (or, equivalently, the relative entropy induced by the word between the distributions over interpretations of the partial sentence) is a determinant of that word's processing difficulty has an interesting and convenient property: surprisal functions as a *causal bottleneck* between representations and behavioral observables.[5] As pointed out in Section 2.1, many different classes of generative stochastic process can determine conditional word probabilities. For any given class of process, there are many different representational choices that may affect the conditional word probabilities that result—including the inventory of states, the independence assumptions between components of the process, and the parameter values that are ultimately chosen or fitted. Under the surprisal (or equivalently, the relative-entropy) theory, however, those representational choices affect predictions about incremental processing difficulty *exclusively* through the conditional word probabilities that they determine. Any two generative stochastic processes that determine the same set of conditional word probabilities will make exactly the same predictions about processing difficulty, regardless of the representational content of these

[5]I am particularly grateful to Andrew Kehler and an anonymous reviewer for helping clarify presentation of the ideas in this section.



(a) Direct effect of representation on processing

(b) Surprisal as a causal bottleneck mediating effect of representation on processing

Figure 1: Surprisal as a causal bottleneck

processes or even the nature of the underlying ("hidden") structures within the process. Furthermore, the probabilistic string model may be more directly inspectable: we might, for example, hypothesize a close relationship between probabilistic string models derived from comprehension models and Cloze probabilities resulting from sentence completion experiments (see Section 4.1 for a more detailed discussion). This property of surprisal contrasts with nearly every other proposed probabilistic theory of sentence comprehension, including competition theories (Section 4.3), the Tuning Hypothesis (Section 4.4), and pruning and attention shift theories (Section 4.5), in which representation affects predictions about processing difficulty more directly. This causal bottleneck property is illustrated schematically in Figure 1b, alongside the contrasting situation in Figure 1a, where representational choices have more direct effects on predictions about comprehension difficulty.

For example, McDonald and Shillcock (2003a,b) show that a word's bigram (also called *transitional*) probability is a significant predictor of reading times for a corpus of eye movements from the reading of British newspaper articles. A bigram word model is a conditional probability model over strings. Referring once again to the causal-bottleneck schematic in Figure 1, it becomes clear that we do *not* need to conclude from McDonald and Shillcock's work that the human parser directly tracks bigrams (although the authors themselves conclude something close to this). We can instead conclude more agnostically that the probabilistic grammatical models used by the human parser for incremental processing and disambiguation determine probabilistic languages that, at a minimum, sensitize the probability of a word w_i to the word w_{i-1} that immediately precedes it. This encompasses a wide range of probabilistic structures, including not only n-grams but also, for example, some types of lexicalized PCFGs (Charniak, 2001). Under the surprisal theory, we might expect that McDonald and Shillcock's results are due to an overall correlation between bigram probabilities and the presumably more refined word surprisals deriving from the human parser's capacity for sophisticated probabilistic disambiguation; and in fact, Frisson et al. (2005) present results suggesting that bigram probability effects on reading times disappear when Cloze probabilities are tightly controlled.[6] The causal bottleneck property is also im-

[6]In addition to the weakly Cloze-controlled study of McDonald and Shillcock (2003a) refuted by Frisson et al. (2005), McDonald and Shillcock also give another reason for concluding that the human parser directly tracks bigrams: they found a significant effect of *backward* transitional probabilities (i.e., $P(w_i|w_{i+1})$) on the reading difficulty of a given word w_i. There is, however, another natural interpretation of this result: the

portant in the analysis of experiments involving German word order in Sections 5 and 7: the precise representation of German word order varies dramatically across different syntactic frameworks, but even relatively simple context-free rules capture the relevant distributional patterns of constituents within German-language sentences. Under surprisal, the finding that reading-time patterns reflect these distributional patterns can be taken as support for the hypothesis that native speakers of German, in the process of online sentence comprehension, construct and are sensitive to statistical information involving descriptions containing information equivalent to these context-free rules.

A closely related point has to do with *bias* in estimating word-by-word comprehension difficulty. Many stochastic string-generating processes (including HMMs and PCFGs) generate unobserved hidden structure "behind" the string whose granularity is not known *a priori*. In order to determine a specific probabilistic model, a granularity level must be chosen and the relevant event probabilities must be estimated with respect to that granularity. Because a word's surprisal is totally dependent on the resulting probabilistic string language, however, a refinement in granularity level will not result in a change in surprisal predictions of a maximum-likelihood estimated model *unless* there are empirical differences in the relevant event probabilities at the finer granularity. As an example, in PCFG modeling we might wonder whether to grammatically distinguish animacy at the level of the noun phrase. Adding a binary animacy distinction to the grammar, for example, would split the following two rules into four:

(1) a. $\text{VP} \rightarrow \text{V NP} \quad \rightsquigarrow \quad \begin{array}{l} \text{VP} \rightarrow \text{V NP[+anim]} \\ \text{VP} \rightarrow \text{V NP[}-\text{anim]} \end{array}$

b. $\text{NP} \rightarrow \text{Det N} \quad \rightsquigarrow \quad \begin{array}{l} \text{NP[+anim]} \rightarrow \text{Det N[+anim]} \\ \text{NP[}-\text{anim]} \rightarrow \text{Det N[}-\text{anim]} \end{array}$

Now suppose that animate and inanimate NPs occur in VPs with the same relative frequency as they occur in the corpus as a whole.[7] Under these circumstances, the resulting animacy-distinguished PCFG still determines exactly the same probabilistic string model, and so its surprisal predictions will be unchanged. (If animate NPs tended to appear disproportionately often or rarely inside VPs, the resulting probabilistic string model and surprisal would of course reflect this, as would be desired.) That is, making the probabilistic grammar more fine-grained has no inherent effect on the predictions about processing difficulty made by surprisal—what matters is whether the finer granularity level leads to the capturing of additional important statistical regularities that the coarser-grained grammar would miss. This lack of granularity-induced bias also contrasts with several other proposed probabilistic theories of syntactic comprehension described in Section 4. For a competition model,

high redundancy of natural English text means that a word's backward transitional probability will generally be correlated with its global left-contextual probability (i.e., $P(w_i|w_{1\cdots i-1})$), even when forward transitional probabilities are accounted for.

[7]Formally, that $\frac{\text{P(VP}\rightarrow\text{ V NP[+anim])}}{\text{P(VP}\rightarrow\text{ V NP[}-\text{anim])}} = \frac{\text{P(NP[+anim])}}{\text{P(NP[}-\text{anim])}}$. If additional rewrite rules for NPs appear in the grammar, then their probabilities for animate and inanimate NPs must also be matched.

for example (Section 4.3), the question arises of whether analyses containing animate and inanimate NP categories compete with each other. For pruning or attention-shift models (Section 4.5), the finer grain size may have an impact on what will get pruned, or what is the top-ranked analysis at a given point.

2.4 Psychological Plausibility

In most cases, a partial input $w_{1\ldots i}$ will be compatible with an infinite number of complete structures T—we can see this simply from the fact that the beginnings of most sentences can be completed in an infinite number of ways. Therefore, it is neither psychologically nor practically possible for the distribution D to be implemented as an enumeration over complete structures. Rather, D would be implicitly determined by some tractable incremental processing algorithm, such as a chart parser (Kay, 1980). Hale (2001) points out that for PCFGs, a probabilistic Earley parser (Stolcke, 1995) determines word-by-word surprisals as a side effect. Other natural formalizations of incremental comprehension might also lead to relative entropy as a natural measure of processing difficulty; the crucial intuition, consistent with the notion of expectation, is that work is required to rule out continuations of an incremental input that might have been, but were not.[8]

The question also arises of how strongly the surprisal theory is a commitment to *full parallelism*: that all possible structural analyses of a sentence are maintained during online comprehension. As Jurafsky (1996) points out, full parallelism becomes less tractable as a wider variety of information sources is brought to bear in probabilistic disambiguation. This leads to the possibility that parallelism in the human parser is *limited*: more than one, but not all, of possible analyses are maintained in the course of online comprehension. Without full parallelism, the strict equivalence between relative entropy and conditional word probability derived in Section 2.1 (and together with it the causal bottleneck property, if one takes the relative entropy measure as primitive) is lost. However, evidence from the probabilistic parsing literature (Roark, 2001; Henderson, 2004) suggests that in typical sentences, most of the probability mass is focused on a small number of highly-ranked analyses. To the extent that this is true, the relative-entropy/surprisal equivalence will be approximate, and the results described in further sections of this paper should remain valid. The results of Sections 6 and 7 include cases where parallelism (maintaining at least two candidates) under surprisal is crucial to giving new explanations for results involving facilitative ambiguity and default grammatical function preferences.

[8] As pointed out by one reviewer, updating a set of probabilities need not in principle be more work- or resource-intensive when the numerical magnitudes of the changes are larger. However, it seems plausible that probabilities allocated to incremental interpretations would be represented by activation levels in relevant structures within the brain, that larger differences in activation levels would correspond to larger physical differences, and that larger physical changes would be more resource-intensive than smaller changes.

3 The structure of probabilistic grammatical models

The goal of this paper is to present an argument for the presence of probabilistically formulated expectation-based effects in syntactic comprehension, and more specifically to advocate a particular relationship—surprisal—between incremental probabilistic disambiguation and processing difficulty. I do not take it as a goal of this paper to advance a particular probabilistic model over trees or strings as the correct one used by adult native speakers of any language. The formulation of such models is the modus operandi of natural language engineering research in parsing and speech recognition, and I take the fact that the best models in these fields today are so sorely outperformed by human capabilities as an indicator that any proposal we can reliably estimate at the present day is almost certainly too simple to be realistic.

Nevertheless, probabilistic grammatical models, in particular probabilistic context-free grammars (PCFGs; Booth 1969), have within applied contexts been remarkably successful in reconciling the tension between broad coverage and ambiguity management that has traditionally plagued computational linguistics (see Collins 1999 and Charniak 1997, among many others). The availability of PCFG models that have the properties 1 and 2 from the beginning of this paper, of robustness to arbitrary input and accurate disambiguation, makes them particularly suitable candidates for estimating surprisal values that can be used to predict patterns in online human comprehension difficulty. For this reason, and also because the experiments analyzed in this paper were constructed to maximize syntactic contrasts, I follow the practice of Hale (2001) in using PCFGs to estimate surprisal values at crucial regions of stimuli in many of the experiments analyzed.[9] In this context, insights gained from applied parsing research can be usefully applied to place minimal requirements on the complexity of accurate probabilistic grammatical models. Natural language parsing research has demonstrated that even from relatively small amounts of data (1 million words or less), the following properties can be reliably incorporated into PCFGs, and in fact *must* be present for accurate disambiguation:

- Gross morphosyntactic properties, such as case marking and agreement features, as well as unbounded syntactic dependencies such as relativization, can be reliably incorporated into the structure of syntactic categories (Collins, 1999; Collins et al., 1999);
- The internal structure of a category may be probabilistically dependent on the lexical (and/or semantic) content of its governor (Magerman, 1994; Collins, 1999; Charniak, 1997);
- Within a local syntactic tree, the distribution of sisters is *history-based*: the presence of a given sister may be probabilistically dependent on which other (both head and

[9] Given a PCFG, existing algorithms by Jelinek and Lafferty (1991) and Stolcke (1995) show us how to calculate the *prefix probability* of a string: the total probability of all trees (or strings) consistent with that prefix. As pointed out by Hale (2001), the conditional probability of w_i is then simply the ratio of the prefix probabilities of $w_{1\cdots i-1}$ and $w_{1\cdots i}$. Precisely which algorithm is used to calculate these probabilities, and details of how the chosen algorithm works, are irrelevant.

non-head) sisters are also present (Collins, 1999; Klein and Manning, 2003);

- The domain of independent events—the probabilistic analogue of the domain of locality in categorical syntactic theory—need not be restricted to local trees (Bod, 1992; Johnson, 1998; Klein and Manning, 2003). For example, the probability that a given NP contains a relative clause can be usefully conditioned on the identity of the NP's parent and sisters.

In all cases where it is practically possible to estimate a complete grammatical model relevant to a particular experiment, PCFG parameters are estimated from a publically available, syntactically annotated corpus of the language in question.[10] In each case, enrichments of the grammatical representation used in the corpus annotation are applied only minimally, to incorporate basic information about the relevant syntactic contrasts in a particular experiment into the grammar, and all enrichments are some subset of the four types listed above. Hence, in all analyses of German clause-final verbs I introduce a categorical distinction between verb-second and verb-final clauses; in the analysis of the effects of varying-size prepositional phrases in Section 5, I distinguish between PPs with three or fewer words from those with four or more; in Sections 5.2 and 7, analyzing the effects of case marking, morphological case is percolated from head nouns onto their NP projections; and in analyzing English relative clauses in Section 5.3, the unbounded syntactic dependency between the relative pronoun and its governing category is threaded through the intervening syntactic categories, in the style of Generalized Phrase Structure Grammar (Gazdar et al., 1985). This minimal approach to grammar refinement ensures that we are not using models more refined than what the rational parser of an adult native speaker could be expected to deploy in probabilistic disambiguation; additionly, minimal grammar refinement reduces both the variance of estimated parameters and the danger of massaging the resulting probabilistic word model to fit observed reading-time patterns. Note, crucially, that while the resulting PCFGs typically have thousands of free parameters, the resulting psycholinguistic model has only one: the amount of difficulty that is caused by one bit of surprisal.

It bears reiterating that this use of PCFGs is not a commitment to any particular grammatical formalism as the backbone of sentence comprehension. PCFGs serve as a formal means of estimating what expectations about upcoming words in a sentence implicitly arise from the use of particular types of information in online sentence comprehension and disambiguation. The choice of a given PCFG embodies a hypothesis about the types of information to which a comprehender's online disambiguation decisions are sensitive, and how that sensitivity is expressed. However, the predictions of the PCFG regarding processing difficulty are completely mediated through the resulting probabilistic word model, as illustrated in Figure 1b. Furthermore, in Sections 6, 7.2, and 8.1, we will not even be able to estimate the parameters of the relevant grammatical model, but by analyzing the conditional word

[10] In all cases, PCFGs are estimated using relative-frequency estimation, with the rewrite of each syntactic category assumed to be probabilistically independent of its ancestors and sisters. See Appendix A for an example of such PCFG estimation.

probabilities in question we will be able to come to firm conclusions about difficulty asymmetries predicted by surprisal under any reasonable probabilistic grammatical model that a native speaker is likely to be using, context-free or not. Finally, this paper make no claims about how much of an adult native speaker's capacity for probabilistic disambiguation is derived from tabulation of statistics directly from individual experience, and how much from higher-order generalizations—innate or learned, linguistic or extra-linguistic—about likely, plausible, or logically possible strings that the comprehender may receive as input. From the perspective of the theory investigated here, the cognitive entity of primary interest is the resulting probabilistic word model alone.

4 Comparison with other processing theories

4.1 Predictability

The surprisal theory bears the greatest conceptual similarity to the well-known observation that words are easier to comprehend in contexts where they are highly predictable (e.g., (2-a) below) than in unconstraining contexts ((2-b)):

(2) a. He mailed the letter without a ***stamp***.
b. There was nothing wrong with the ***car***.

This effect of predictability has been observed in both eye-tracking reading studies, as reduced reading time and increased skipping probability (e.g., Ehrlich and Rayner 1981), and in evoked-reaction potential (ERP) studies, as a differential N400 effect (e.g., Kutas and Hillyard 1980, 1984). The traditional method of quantifying predictability has been the use of Cloze completion studies (Taylor, 1953), where the predictability is measured as the probability with which subjects complete an initial context such as *he mailed the letter without a* __ with the word of interest, such as ***stamp*** in (2-a) above.

In expectation-based theories as formulated here and in Hale (2001), the crucial measure of surprisal is conceptually something very close to a negative log-Cloze probability, and indeed the surprisal of an extremely predictable word should be lower than a somewhat predictable word (such as a Cloze probability differential of 0.9 versus 0.6; Kutas and Hillyard 1984; Federmeier and Kutas 1999). The surprisal theory goes beyond the traditional domain of predictability in three respects, however. First, the theory proposes that conditional probability affects difficulty in a *log scale*. That is, the ratio rather than the difference between the conditional (or Cloze) probabilities should be the determinant of differential difficulty between two items; so we should see similar effects between items with probabilities 0.05 and 0.1 as between items with probabilities 0.5 and 1. Recent modeling literature on predictability effects in reading has assumed that they function on an absolute probability scale (Reichle et al., 1998; Rayner et al., 2004; Engbert et al., 2005), but the results of Rayner and Well (1996) suggest that similar absolute differences in predictability have a greater impact on difficulty on the low end of the scale than on the high end of the scale, which is

expected under the surprisal theory.[11] Second, the surprisal theory explicitly predicts that we should see differential predictability effects even for words that are not the most likely completion of a given context. Third, predictability is generally considered a primarily semantic phenomenon; but surprisal differences can derive from any source, including syntax, morphology and phonology as well as semantics. The bulk of studies examined in the remainder of this paper involve differences in (quite small) conditional probabilities deriving from syntactic effects; in many cases, the relevant objects of prediction can be thought of as syntactic categories rather than wordforms.

4.2 Locality

Locality-based processing theories include two hallmark proposals. The first is that greater distance between entities in a syntactic relationship causes greater difficulty when that relationship is constructed; Gibson's Dependency Locality Theory (DLT; Gibson 1998, 2000) is an exemplar of this type of proposal. Under the DLT, the comprehension difficulty of a word w is taken to be affected by (among other factors) its structural integration cost, which is monotonically increasing in (a) the number of dependency relationships between w and words that precede it; and (b) the distances between w and the preceding words with which it is in a dependency relation. The second hallmark proposal is that preference for more local syntactic relationships directly guides disambiguation, and when maximally local structures turn out to be wrong, difficulty is incurred because the parser has been misled. This proposal has had a wider variety of incarnations; perhaps the most prominent current incarnation is the Active Filler Hypothesis (AFH; Clifton and Frazier 1989).

Head-final local syntactic dependencies turn out to be a rich source of divergence between predictions of the DLT and surprisal. There are a variety of syntactic circumstances in which a comprehender knows that a final governing category has to appear, but does not know exactly when it will appear, or what it will be. This situation is common in languages with obligatorally verb-final clauses, such as in German, Japanese, or Hindi. As Konieczny (2000) points out, the DLT predicts in these cases that a larger number of left dependents will cause greater processing difficulty at the final governor, because all the left dependents must be integrated with it at the same time. But the surprisal theory makes the *opposite* prediction in this case. The more dependents we have seen, the more information we have about their governor, and in general the more information we have, the more accurately we should be able to predict that governor's location and identity.[12] Experiments testing this

[11] In addition, word frequency is generally taken to affect difficulty on a log scale. Because the surprisal theory as derived in Section 2 is a consequence of generative probabilistic models, it subsumes word frequency as a part of conditional word probability. Word frequency corresponds simply to a unigram probabilistic word model.

[12] Konieczny informally makes a similar point: extra dependents can help us narrow down the class of events that a final verb might denote, and therefore aid in lexical access. The surprisal theory encompasses this position, which involves prediction of the *identity* of the item ending the clause, but is more general, as it includes predictions about the *location* of the end of the clause. See discussion of Jaeger et al. (2005) in Section 5.3 for evidence that humans make accurate, syntactically-driven positional predictions consistent with the surprisal theory.

divergence in prediction have been carried out by Konieczny (2000); Konieczny and Döring (2003); Vasishth and Lewis (2006); Gibson et al. (2005b) that are informally consistent with surprisal's predictions. In Sections 5.1 and 5.2 I construct explicit expectation-based models of German verb-final clauses, showing that predictions of the surprisal theory closely match qualitative reading-time patterns.

DLT-style and AFH-style theories make similar predictions regarding long-distance dependencies that violate minimal locality, such as object over subject relativizations (e.g., King and Just 1991; Gibson et al. 2005a):

(3) a. The reporter *who* attacked the senator admitted the error.
b. The reporter *who* the senator attacked admitted the error. (Gibson, 1998)

In the DLT, the object extraction (3-b) is predicted to be more difficult than the subject extraction (3-a), due to the storage cost of maintaining the extracted element longer plus the final cost of a longer-distance integration. In the AFH, the parser greedily posits a gap immediately to match the relative pronoun filler; when that decision turns out to be incorrect, as in (3-b), reanalysis is required and difficulty ensues.

Surprisal-based processing predicts the same general asymmetric difficulty, but for a different reason: in the above examples, extractions from the leftmost site are more common than more distant extractions. Hale (2001) showed that the simplest PCFGs derived from annotated corpora of English text assign a higher surprisal to object-extracted relative clauses such as (3-b) than to subject-extracted relative clauses, essentially because most relative clauses are subject extractions. The predictions of the DLT, the AFH, and surprisal begin to diverge, however, at the finer-grained level of exactly *where* processing difficulty is predicted to occur in nonlocal dependencies, and subsequent sections of this paper analyzes several relevant experiments. Section 7 presents a detailed analysis of AFH-inspired experiments from Schlesewsky et al. (2000) on German verb-second clauses that the authors contend undermine serial frequency-based processing accounts, and shows that surprisal actually models these experiments more precisely than the AFH itself. Section 8.1 presents analysis of two detailed recent studies on English relative clauses (Gordon et al., 2004; Grodner and Gibson, 2005), which yield some results consistent with surprisal but also seem to support a locality-based component of syntactic processing difficulty.

4.3 Competition and dynamical models

Traditionally, parallel constraint-satisfaction models of syntactic comprehension have taken *competition* as the link connecting incremental disambiguation and observable measures of processing difficulty (MacDonald et al., 1994; Spivey and Tanenhaus, 1998; McRae et al., 1998). In these models, a variety of noncategorical, weighted constraints, potentially extralinguistic as well as linguistic, are simultaneously brought to bear in the incremental disambiguation of syntactic ambiguity. In these dynamical models, candidate analyses of an input substring compete with each other to reach a critical activation threshold, and the number of cycles in the network that it takes to reach this threshold determines predicted

processing times. As a result, the greater the total weight of constraints satisfied by the favored analysis relative to alternate analyses, the faster this analysis can reach activation and the easier comprehension will be. This leads to at least two types of empirically predicted high reading times. First, if an early part of the input causes one analysis to be favored, but later parts of the input disconfirm that analysis in favor of another, it can take time for the system to gravitate from the original to the new analysis. (This can been seen as a form of attention shift, as discussed in Section 4.5.) Second, when the system is near a boundary between multiple analyses, it can linger in a state of competitive gridlock.[13] Competition model can be thought of as resource-allocation models in which a fixed amount of resources must be distributed among competing analyses of a partial input, and in which (unlike for surprisal) allocation leading to relative equibias among analyses is inefficient.

It merits notice that competition models proposed in the literature contain two logically separable components: the integration of multiple, non-categorical constraints as the mode of syntactic disambiguation; and the attribution of long processing times to competition among alternative interpretations. The surprisal theory is completely compatible with the first component, but not with the second: under surprisal, difficulty occurs when resources are distributed in a way that is not highly compatible with the continuation of the sentence. Differences between the predictions of surprisal and competition models will become clear in Section 6, where experiments are analyzed that may demonstrate circumstances under which unresolved ambiguity can speed comprehension; and in Section 8.3, in which experiments seem to indicate that under some circumstances, purely locally coherent syntactic analyses can compete with global probabilistic expectations.

4.4 Tuning

The tuning hypothesis (Mitchell, 1994; Cuetos et al., 1996; Mitchell et al., 1995) is a serial-choice model of syntactic disambiguation in which syntactic ambiguity is resolved by choosing the most frequent structural variant. The processing difficulty that ensues when a subsequent word is consistent with only the less frequent variant is considered a mild form of garden pathing. The surprisal theory agrees with the tuning hypothesis in assuming the rationality hypothesis that more frequent structural variants are preferred, but differs in its commitment to parallelism and in the formalization of processing difficulty. Nevertheless, for the head-initial structures that have been of primary interest in tuning hypothesis research—such as leftward attachment of relative clauses into multilevel NPs—the predictions of surprisal are essentially similar to those of tuning.[14] As will be seen in Section 5, however, for head-final structures the surprisal theory makes substantial predictions in cases where the tuning hypothesis has nothing to say. The tuning hypothesis is also sensitive to granularity bias, as discussed in Section 2.2. Finally, Section 6 discusses differences in how serial-choice and parallel surprisal theories deal with cases where unresolved ambiguity can facilitate comprehension.

[13] See also the analysis of Tabor and Tanenhaus (1999), who show how competition effects can also emerge directly from a predictively trained neural network.

[14] This follows directly from the Markov decomposition of conditional word probability, plus Bayes' rule:

4.5 Pruning and Attention Shift

A number of other ranked-parallel syntactic comprehension models have been proposed (Gibson, 1991; Jurafsky, 1996; Narayanan and Jurafsky, 1998, 2002; Crocker and Brants, 2000); as in surprisal, comprehension difficulty in these models is a function of the rankings of possible structure before and after a given word. The most prominent sources of difficulty in these models have been *pruning*, where low-ranked structures can be eliminated due to memory limitations, and *attention shift*, where change in what the highest-ranked structure is causes difficulty. These models of processing difficulty are thus immediately distinguished from surprisal in that there are no known causal bottlenecks for these theories without the introduction of additional specialized assumptions, since it is impossible in general to determine whether a particular structure is the highest-ranked or has been dropped altogether without making a specification of what the set of possible structures actually is. Nevertheless, the surprisal theory captures key insights involved in both attention shift and pruning. Consider, for example, a situation with two major incremental interpretations I_1 and I_2, at a point where I_1 has a conditional probability well over 0.5. A word that causes I_2 to become most probable will necessarily involve considerable surprisal: this extra surprisal corresponds to the attention shift effect proposed in Narayanan and Jurafsky (2002).[15] Likewise, when a given word w can only be generated from an unlikely structure S, the conditional probability of w can be no higher (and will typically be much lower) than the conditional probability of S, as in classic garden-path sentences: as originally demonstrated by Hale (2001), this gives an effect quite similar to that of pruning models. In the surprisal theory, however, attention-shift and pruning effects are not all-or-nothing, and are thus compatible with difficulty gradients such as those demonstrated for different types of reduced relative clauses in work such as MacDonald et al. (1994) and Spivey and Tanenhaus (1998).

$$\begin{aligned} P_i(w) &= \sum_T P(T|w_{1\cdots i-1})P(w|T) \\ &= \sum_T \frac{P(T, w_{1\cdots i-1})}{P(w_{1\cdots i-1})}P(w|T) \\ &\propto \sum_T P(T)P(w|T) \end{aligned}$$

where T range over the possible *partial* parses of $w_{1\cdots i-1}$. When only one partial parse T^* can generate w, the conditional word probability $P_i(w)$ will be larger when the structural frequency, and hence probability, of T^* is greater, assuming that $P(w|T)$ is approximately constant (which, generally, is implicitly ensured in the relevant work by controlling for factors such as plausibility and word frequency at the disambiguating word).

[15] Using the relative-entropy derivation, we can actually put a (conservative) lower bound on the surprisal from the contribution of the I_2 term alone: $P_{k+1}(I_2) \log \frac{P_{k+1}(I_2)}{P_k(I_1)}$.

4.6 Prediction-based connectionist models

Many connectionist models of online sentence processing (including Elman 1990, 1991; Christiansen and Chater 1999; Tabor and Tanenhaus 1999; Konieczny and Döring 2003; and a component of the model of Rohde 2002) propose *prediction-based* metrics of the difficulty of a word w_i in its sentence context with the activation of w_i after $w_{1\ldots i-1}$ have been seen—the lower the activation, the greater the difficulty. The activation at the output layer of a predictive neural net can be directly interpreted as a multinomial probability distribution over the next input token, and the most commonly used training regimens can be seen as directly optimizing the predictive power of the net (see Rumelhart et al. 1995, inter alia for discussion). Depending on the precise definition used, these difficulty metrics can be equivalent to or quite similar to the surprisal metric proposed here.[16] As a result, some predictive connectionist models of online sentence comprehension may make predictions about reading times quite similar to those presented in the paper. The precise predictions can vary significantly, of course, based on the model underlying conditional word probabilities—this paper emphasizes the use of hierarchically-structured probabilistic grammars estimated from syntactically annotated corpora, whereas most connectionist models are trained on corpora consisting of raw text (that is, word strings).[17] In some cases, researchers using connectionist models have drawn a strong link between the process of grammar acquisition and results in adult native-speaker sentence processing. Christiansen and Chater (1999), for example, have argued that structural relationships within strings that are hard for networks to learn, such as nested dependencies, are also the hardest for adult native speakers to process. [18] The theory proposed here, in contrast, assumes that the structural relationships underlying surface strings are learned perfectly, similarly to other proposals discussed earlier in this section.

5 Verb-final contexts, surprisal, and locality

There are contexts in nearly every language where a head follows one or more of its dependents. When a language comprehender recognizes that a partial input has entered such a context, they are in a position where they obtain increasing amounts of information about the upcoming head. Intuitively, this accumulating information has two effects: on the one hand it places a greater memory load on the comprehender, on the other hand it can help sharpen comprehenders' expectations about the upcoming head. This situation is perhaps most ubiquitous in languages where verbs are final in their clause, such as German (excluding finite matrix-clause verbs), Japanese, and Hindi. Unlike in English, when a clause-final

[16] I am grateful to an anonymous reviewer for this point.

[17] Of particular interest in this connection is the model of Henderson (2004), in which a probabilistic model over context-free trees is learned, but a neural net is used to learn a compressed yet potentially unbounded history representation over conditioning structure.

[18] Other connectionist models of online sentence processing, such as Tabor et al. (1997) and Tabor and Tanenhaus (1999), propose metrics of online processing difficulty that are not based on prediction, but retain the tight link between connectionist acquisition and online processing.

verb is encountered the number and distance of previous dependents can vary widely. As pointed out by Konieczny (2000), DLT-type locality theories predict that the final verb will be more difficult to process when it has a greater number of dependents. Section 4.2 argues informally that surprisal predicts the opposite: more preverbal dependents gives the comprehender more information with which to predict the final verb's identity and location, and comprehension should therefore be easier.[19] In the last several years, a number of reading studies have been reported which bear upon this divergence in predictions. Sections 5.1 and 5.2 presents a surprisal-based analysis of Konieczny (2000) and Konieczny and Döring (2003), for which the resources exist to construct explicit computational surprisal-based models. Section 5.3 analyzes another upcoming-head experiment, this time in English, and Section 5.4 briefly discusses related experiments in Hindi and Japanese (Vasishth and Lewis, 2006; Gibson et al., 2005b; Nakatani and Gibson, 2003).

5.1 Konieczny 2000: effect of additional constituents

Konieczny (2000) was the first to investigate the effect of extra preverbal constituents on processing difficulty, measuring reading time at clause-final verb in transitive German embedded clauses where the amount and type of material between the direct object and the final verb varied, as in (4) below.

(4) a. Er hat den Abgeordneten begleitet, und ...
He has the delegate escorted, and ...
"He escorted the delegate, and ..."

b. Er hat den Abgeordneten ans Rednerpult begleitet, und ...
He has the delegate to_the lectern escorted, and ...
"He escorted the delegate to the lectern, and ..."

c. Er hat den Abgeordneten an das große Rednerpult begleitet, und ...
He has the delegate to the big lectern escorted, and ...
"He escorted the delegate to the large lectern, and ..."

In (4-a) the verb directly follows the direct object; in (4-b)-(4-c) a prepositional phrase goal of varying size intervenes between the direct object and the verb. From a locality-based perspective the predictions are clear: the verb should be easiest to process in (4-a), because it has the fewest and nearest dependents; and hardest to process in (4-c), because it has the most and farthest dependents. Konieczny, however, found the opposite pattern: the verb was processed the fastest in (4-c) and slowest in (4-a) (see Table 1).

In order to determine the predictions of surprisal-based sentence processing on Konieczny's data, it is necessary to choose a probabilistic language model $p_i(w)$. The choice of model should be driven by our linking hypothesis between incremental comprehension and difficulty: the model chosen as optimal for purposes of incremental processing and disambiguation should accurately predict per-word reading times. In this case, our data—the

[19] Assuming, of course, that the identity of the final verb is consistent with the contents of its preverbal dependents.

	Average RT (ms)	Surprisal	DLT prediction
no PP	514	15.99	faster
short PP	477	15.41	slower
long PP	463	15.35	slower

Table 1: Empirical reading time versus surprisal at clause-final verb of (4)

experimental stimuli used in reading-time experiments—do not follow ecologically natural distributions, but rather maximize clause-level structural variation while minimizing other structural and lexical variation. A non-lexicalized PCFG is therefore a sufficient basis for modeling the contrast observed in Konieczny's data. We can take advantage of the hand-parsed NEGRA corpus (Skut et al., 1997) of German, and use essentially the grammar read straight off the parsed corpus to construct a language model, making only minimal changes to the grammatical representations in the corpus necessary to encode important distributional properties of German syntax not already directly encoded in the local-tree structure of the NEGRA corpus.[20] Calculating the final-verb surprisals of one of Konieczny's items, given in (4), and comparing it to reported mean reading times results in the comparison shown in Table 1. As can be seen, surprisal values match average reading times quite closely. The DLT, in contrast, predicts the wrong monotonicity of reading difficulty.[21]

The reason that PCFG-derived surprisal values match Konieczny's empirical results so well is that incremental parsing with a PCFG naturally captures the effect of a sentence's *constituent history* on the expectations regarding yet-to-be-seen input. As soon as the comprehender knows that the input is part of a verb-final clause, the incremental probabilistic parsing process implicitly determines set of expectations as to the next constituent. Each subsequent constituent affects these expectations. To a first approximation, seeing a constituent of a given type (a subject, a direct object, the final verb, a goal, a location, and so on) sharply decreases the expectation of seeing another constituent of the same type in the same clause, because multiple constituents of a single type rarely co-occur in a single clause; this is part of the comprehender's knowledge of linguistic argument structure, captured in the PCFG model by the structure of rewrite rules. When a PP goal is actually seen in

[20]For all models of German-language experiments, in order to sharpen the PCFG's distributional knowledge of V2 versus verb-final contexts I introduced syntactic distinctions in the VP and CVP (coordinated VP) NEGRA syntactic categories based on whether the clause was matrix or subordinate. Subordinate-clause VPs were defined as those under an S category and sister to a PRELS tag, which is the NEGRA syntactic category for relative pronouns. For the model of Konieczny (2000), I additionally follow his particular experimental design to distinguish *small* PPs (those with 2-3 words) and *large* PPs (4-6 words) from other PPs. This last distinction is motivated by the fact that variable word-order phenomena such as heavy NP shift (Wasow, 2002) and right-extraposition (Uszkoreit et al., 1998) are highly sensitive to constituent size.

[21]Konieczny used a variety of experimental items, the word-by-word surprisals of most of which could not be calculated due to lack of lexical coverage in the NEGRA corpus. For every item that was covered by NEGRA, the monotonicity of surprisal is the same in the pairwise contrasts between presence and absence of PPs. The small/large PP contrast had correct monotonicity in only half the items, but the mean surprisal difference was 0.50 bits in the correct direction.

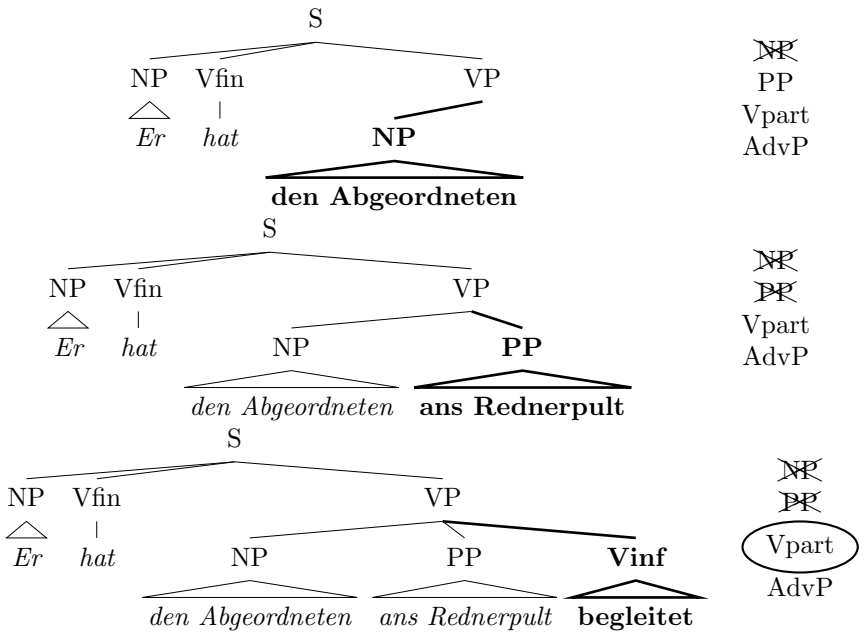


Figure 2: Incremental parse of (4), showing incremental narrowing of next-constituent syntactic expectations

the input, as in (4-b), the expectation allocated to seeing a PP goal is pruned away, and because expectation is actually a probability distribution that must sum to 1 at all times, it is reallocated among all the other types of constituents that have not yet been seen. The final verb, being one of those constituents, therefore has its expectation increased after every other constituent. In another manner of speaking, the comprehender's expectation as to the *location* of the final verb sharpens as the clause lengthens. The way this incremental expectation-narrowing process plays out in a PCFG-derived probabilistic string model is illustrated in Figure 2: as each constituent of a given category is seen and integrated into the incremental parse, it eliminates most of the expectation for seeing another constituent of the same type next, and as a result increases the expectation for seeing a constituent of one of the remaining types.[22]

5.2 Konieczny and Döring 2003: effect of preverbal NP type

Konieczny and Döring (2003) report a variant of Konieczny (2000)'s original experiment, where the syntactic position of a preverbal NP, rather than the presence/absence of a preverbal PP, is varied:

[22]Technically, the PCFG used to model this experiment does not distinguish goal PPs from other types of PPs, because the NEGRA corpus unfortunately does not make this distinction. The PP category in Figure 2 is therefore not subdivided. Nevertheless, the same intuitive argument holds for this cruder grammatical model, because PPs in general are in complementary distribution with each other in verb-final contexts.

(5) a. Die Einsicht, daß [$_{\text{NP}_{\text{NOM}}}$ der Freund] [$_{\text{NP}_{\text{DAT}}}$ dem Kunden] [$_{\text{NP}_{\text{ACC}}}$ das Auto
the insight, that the friend the client the car
aus Plastik] verkaufte, ...
from plastic sold, ...
"The insight that the friend sold the client the plastic car ..."

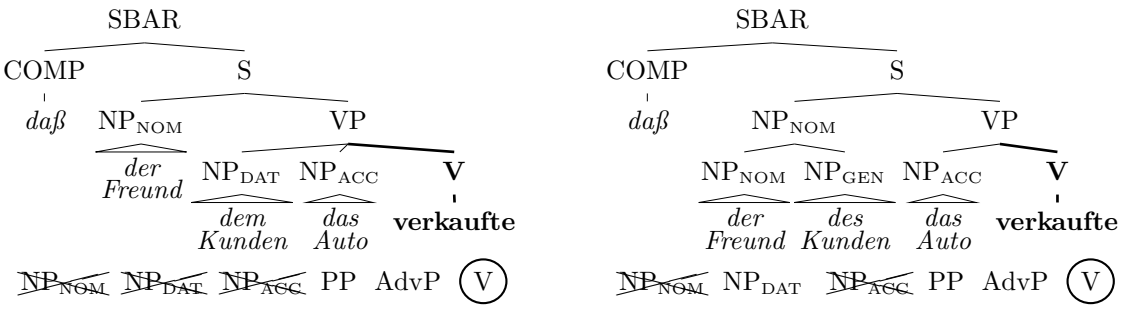
b. Die Einsicht, daß [$_{\text{NP}_{\text{NOM}}}$ der Freund [$_{\text{NP}_{\text{GEN}}}$ des Kunden]] [$_{\text{NP}_{\text{ACC}}}$ das Auto aus
the insight, that the friend the client the car from
Plastik] verkaufte, ...
plastic sold, ...
"The insight that the friend of the client sold the plastic car ..."

In an eye-tracking reading study, Konieczny and Döring found that regression-path times for the final verb *verkaufte* were significantly shorter for the dative condition, where *dem Kunden* is dependent on the final verb, than for the genitive condition, where *des Kundes* is dependent on the preceding noun *Freund*.[23] This study is a nice methodological confirmation of the original pattern observed in Konieczny (2000). The stimuli in (5) differ in only a single letter, thus controlling quite precisely for the orthographic length, number of tokens, and also, as it turns out, word frequency of the material preceding the critical region (*dem* and *des* are quite close in overall word frequency, with *des* perhaps slightly more frequent in some contexts).

Intuitively, surprisal applies just as readily to this experiment as to Konieczny's original experiment. Just before seeing the final verb in (5-a), the comprehender knows that nominative, accusative, and dative NP arguments have all appeared as preverbal dependents; in (5-b), only nominative and accusative preverbal dependents have appeared. The comprehender's expectations are therefore more narrowly focused in (5-a), and so the surprisal at the final verb should be lower. In order to precisely model this effect of constituent history, we can use a PCFG grammar to determine a conditional word model as we did in Section 5. Here, however, the crucial difference in experimental conditions involves case marking on an NP constituent. Fortunately, about a third of the NEGRA corpus includes case-marking annotation on the wordforms, and we can transfer this information up to phrasal nodes using simple grammatical rules, so that the learned PCFG captures basic distributional generalizations about case-marking patterns in German.[24] Knowledge of the overall distribution of *argument realization frames*, such as the rarity of multiple dative NPs in a single clause, is thereby transferred into the PCFG. We then use these case-enriched symbols as atomic categories, and learn a PCFG via relative-frequency estimation from the enriched corpus. The

[23]They also varied whether the immediately preverbal PP was a nominal dependent, as in *aus Plastik* in (5), or a verbal dependent such as *aus Freude*. Although they found slightly shorter average reading time for the nominal-dependent case, this difference was not statistically significant. If subsequent studies were to achieve a statistically significant result favoring faster reading times for the nominal-dependent condition, then it could be problematic for the expectation-based account presented here.

[24]To be precise, we recursively percolate case marking onto NPs and PPs from their head daughters (the case of a preposition is considered to be the case it governs). These percolation rules are similar in nature to constraints used in unification-based grammatical formalisms such as Functional Unification Grammar, Head-Driven Phrase Structure Grammar, and Lexical-Functional Grammar.



(a) Verbal-dependent condition (b) Nominal-dependent condition

Figure 3: Incremental parsing with case-percolated PCFG

	Reading time (ms)	surprisal	DLT prediction
verbal dependent (dative)	555	23.51	slower
nominal dependent (genitive)	793	23.91	faster

Table 2: Reading time, surprisal, and DLT predictions at final verb for (5)

incremental parses of (5) are shown in Figure 3, together with schematics of next-constituent expectations at the point of seeing the final verb.

As we saw previously in Figure 2 for Konieczny (2000), the extra preverbal constituent in the verbal-depend condition sharpens next-constituent expectations and thereby decreases surprisal at the final verb itself. Table 2 shows empirical regression-path reading times, conditional word probabilities, and DLT-predicted reading times for the two conditions of (5). Although the conditional probability of the final verb is quite low in both conditions, it is roughly 30% higher in the verbal-dependent condition than in the nominal-dependent condition, correctly predicting reading time monotonicity. DLT, on the other hand, predicts faster reading time for the nominal-dependent condition, since there are fewer preverbal dependents for the verb to integrate with.

5.3 Disentangling verb location from verb identity

The experimental results described in Sections 5.1 and 5.2 are also compatible with the informal intuition of Konieczny (2000) (see also Konieczny 1996), that preverbal dependents constrain the lexical type of the final verb and thus allow better prediction of that verb. The surprisal analysis, however, shows that this explanation based on verb *identity* is not strictly necessary. Assuming only an unlexicalized PCFG—that is, assuming that native German speakers are capable of discriminating good from bad constituency structures—determines a surprisal model reflecting only information about verb *location* that nevertheless makes qualitatively correct predictions about final verb reading times.

This section shows how information about verb location and identity can be disentangled, by discussing a recent experiment carried out by Jaeger et al. (2005). Jaeger et al. used

English subject-modifying relative clauses of varying lengths, and observed reading times occurring on the matrix-clause verbs appearing immediately after the RC:

(6) a. The player [that the coach met **at 8 o'clock**] bought the house...
b. The player [that the coach met *by the river* **at 8 o'clock**] bought the house...
c. The player [that the coach met NEAR THE GYM *by the river* **at 8 o'clock**] bought the house...

Like German verb-final clauses, English subject-modifying relative clauses are a constrained syntactic context. The comprehender knows that the relative clause has to end, but does not know when it will end until seeing the next item of the matrix clause (in this case, the matrix verb). The more postverbal constituents within the RC that have been seen, the fewer possible choices there are for subsequent constituents within the RC. This follows because constituent types tend to be in complementary distribution—for example, in a given clause the knowledge that a temporal phrase has already appeared makes it less likely that a new temporal phrase will be seen. This means that the comprehender's expectation for the end of the RC (and hence seeing the matrix verb next) should generally increase as the number of already-seen postverbal constituents increases. The DLT, in contrast, predicts that more RC-internal constituents will lead to greater matrix-verb difficulty, as the distance from the matrix subject it governs increases. The predictions of the surprisal theory can be made precise by using an unlexicalized PCFG of English, learned from the parsed Brown corpus section of the Penn Treebank (Marcus et al., 1994). The Brown corpus represents multiple postverbal dependents as sisters within a single local tree, meaning that the resulting PCFG encodes the relevant distributional dependencies among postverbal dependents.

Table 3 shows matrix-verb surprisal values estimated by a PCFG trained directly off the parsed Brown corpus, together with DLT predictions and empirical mean reading times.[25] The surprisal model matches empirical results: surprisal and reading time at the matrix verb both decrease as the number of postverbal constituents in the preceding RC increases. Crucially, the observed effect does *not* follow from the account of Konieczny (1996, 2000), in which preverbal dependents help the comprehender guess the *identity* of the final verb, because there is no direct argument structure relation between the matrix verb and the verbal dependents in the RC. This effect is also unpredicted by a theory of anti-locality effects proposed Vasishth and Lewis (2006), under which a governing head can be primed by preceding constituents that are (i) its dependents, or (ii) dependents of its dependents; the extra PPs in (6) are neither. The broader surprisal theory encompasses the narrowing of expectations proposed by Konieczny for final-verb identities, but also predicts that comprehension patterns will reflect implicitly-formed expectations about upcoming constituency, a prediction that is borne out in this experiment.

[25]The paired comparisons between the 1 and 2 and 1 and 3 PP conditions are statistically significant; the paired comparison between the 2 and 3 conditions is not.

	Number of PPs intervening between embedded and matrix verb		
	1 PP	2 PPs	3 PPs
DLT prediction	Easier	Harder	Hardest
Surprisal	13.87	13.54	13.40
Mean Reading Time (ms)	510 ± 34	410 ± 21	394 ± 16

Table 3: Surprisal and average reading times at matrix verb for (6)

5.4 Other investigations involving verb-constraining contexts

Reading-time investigations of clause-final verbs have also been carried out in Hindi and Japanese. (Vasishth 2002 (Chapter 5), 2003, Vasishth and Lewis 2006) have conducted several experiments on processing difficulty within Hindi complement clauses and relative clauses, both of which are verb-final, varying the amount of material appearing before the final verb. Consistent with predictive accounts including surprisal, reading time at final embedded verbs is lowest when there is more preverbal material within the clause.

Two other relevant experiments have been carried out by Nakatani and Gibson (2003) and Gibson et al. (2005b) for Japanese, which is verb-final and has freely reorderable preverbal complements. In both cases, predicted asymmetries in integration cost at final verbs failed to emerge. Gibson et al. (2005b) found patterns similar to those we have already seen in German and Hindi: greater amounts of preverbal material decreased, rather than increased, final-verb reading times.[26] For Nakatani and Gibson (2003), the object of investigation was the degree of center-embeddedness in sentences with multiple sentential complements. They found that the greatest difficulty associated with multiply center-embedded sentences occurred at the onset of the most deeply embedded clause—signaled by a third consecutive animate nominative NP at the beginning of the sentence—rather than at the final, least-embedded main verb, where the integration-cost component of DLT predicts it to occur. In the surprisal theory, the natural place to look for an explanation of this result would be to estimate the probability of the conditional probability of a third consecutive animate, nominative NP given two such consecutive sentence-initial NPs. Unfortunately, large annotated corpora of Hindi and Japanese are not readily available, so more detailed and explicit models addressing these issues must remain as topics of future research.

5.5 Discussion

Explicit word-probability models constructed using PCFGs trained on hand-annotated corpora of German provide a qualitative match to empirical reading-time differences found at

[26] Unlike the results of Konieczny and Döring (2003), Gibson et al. found no difference in the reading times in contrasts of adverbial versus adnominal positioning of a preverbal constituent. One plausible explanation could be that whereas the verbal/nominal dependency alternation in Konieczny and Döring was confounded with the ditransitivity/ditransitivity alternation, in Gibson et al. (2005b) it involved a locative constituent, which may have facilitated similar evidential inferences about the final verb from either position.

clause-final verbs. Because these models are unlexicalized, their predictions reflect only the incremental change into comprehenders' expectations about the *location* of the yet-unseen final verb. These results demonstrate that, under the surprisal theory, even quite simple models of probabilistic ambiguity resolution naturally give rise to the prevalent pattern of reading-time results observed in verb-final clauses: additional dependents facilitate comprehension of the final verb.

In a probabilistic language model that closely matched naturally-occurring corpus data, however, additional preverbal dependents would also sharpen expectations regarding the verb's *identity*. As Konieczny (1996, 2000) himself points out, seeing a goal PP restricts the syntactic/semantic classes from which the final verb can plausibly originate.[27] Konieczny and Döring present a constraint-based computational model in the form of a simple recurrent network (SRN; Elman 1990) that captures probabilistic dependencies between specific verbs and their host of dependents. Their SRN, trained on an artificially generated corpus consisting of both transitive and ditransitive verbs, models *probabilistic lexical selection preferences*: the presence of a preverbal dative argument excludes simple transitive final verbs from the space of possible final verbs, and hence boosts the expectation for those ditransitive verbs the dative argument has been seen to occur with. They do not formalize a general relationship between these expectations and reading time predictions, but their SRN's results match their experimental results under any theory in which a word's processing difficulty decreases monotonically with the activation its output node in an SRN (including under surprisal, if output node activation levels are interpreted as a probability distribution over the next word, as in Rumelhart et al., 1995). In principle, lexical selectional preferences and expectations about verb location could be combined in a surprisal model using lexicalized PCFGs (see Collins 1999) trained on naturally occurring German data, rather than an artificial corpus. Unfortunately, corpus data is too sparse to easily yield reliable estimates of lexical selectional preferences for all but the most common verbs; *verkaufte* 'sold', for example, occurs only five times in the NEGRA corpus (see also Dubey and Keller 2003). As we have seen, however, even unlexicalized PCFG surprisal models predict the correct monotonicity of difficulty not only for clause-final verbs, where information abut verb identity and verb location are conflated, but also for English matrix-clause verbs as in Section 5.3, where only information about verb location is likely to be relevant.

6 When ambiguity facilitates comprehension

The fully-parallel surprisal theory entails an unusual relationship between structural ambiguity and processing difficulty. In most processing theories, local structural ambiguity leads to difficulty under a variety of circumstances. In serial theories, local ambiguity is a precondition for garden-path effects; in competition-based parallel accounts, equibias while an ambiguity is unresolved is the primary source of syntactic comprehension difficulty. In

[27] Vasishth and Lewis (2006) propose an account based on activation decay and retrieval inference (Anderson et al., 2004) similar in many respects to that of Konieczny (1996), except that the preverbal constituents prime related final verb candidates rather than rule out incompatible candidates.

the surprisal theory, on the other hand, structural ambiguity per se plays no role in the determination of processing difficulty: ambiguities are relevant only insofar as they have an effect on conditional word probabilities. In the language of probability theory, a word w_i's surprisal *marginalizes* over all the possible partial structural descriptions consistent with the string prefix $w_{1\cdots i}$:

$$P_i(w) = \sum_{T, w_{1\cdots i-1} \in T} P(T|w_{1\cdots i-1}) P(w_i|T)$$

where T ranges over the partial syntactic/semantic structures of $w_{1\cdots i-1}$. If there is a local ambiguity through $w_{1\cdots i-1}$, and more than one structural variant T can give rise to a given next word w, then the conditional probability $P_i(w)$ draws mass from all these T; the multiple variants can be said to *conspire* to facilitate processing of w. This is in sharp contrast to competition-based accounts, in which the multiple possible variants compete with each other and thus should impede processing of w.

This direct prediction of the surprisal theory turns out to bear directly on a number of established findings. Most prominent are the results of Traxler et al. (1998); van Gompel et al. (2001, 2005), who show that ambiguous left attachments into a complex NP are, if anything, read more *quickly* when they do not resolve the attachment level. Example (7) below illustrates a characteristic finding in this work.

(7) (Traxler et al., 1998)

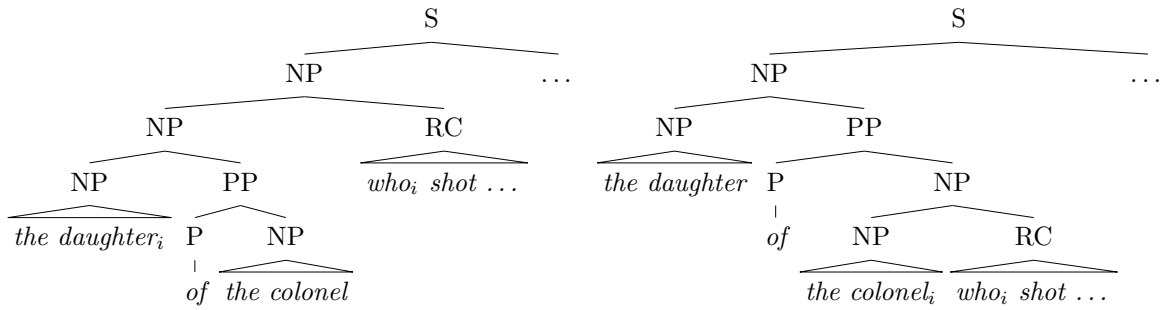
a. The daughter$_{\text{i}}$ of the colonel$_{\text{j}}$ who shot herself$_{\text{i/*j}}$ on the balcony had been very depressed.

b. The daughter$_{\text{i}}$ of the colonel$_{\text{j}}$ who shot himself$_{\text{*i/j}}$ on the balcony had been very depressed.

c. The son$_{\text{i}}$ of the colonel$_{\text{j}}$ who shot himself$_{\text{i/j}}$ on the balcony had been very depressed.

In (7-a) and (7-b), the reflexive pronoun disambiguates the locally ambiguous attachment of the relative clause *who shot...*; in (7-c), the reflexive pronoun has ambiguous antecedence, and both high and low attachment are possible. Traxler et al. found that the reflexive pronoun and surrounding regions were read more quickly and with fewer regressions in the ambiguous case than in either ambiguous case. Analogous reading patterns have been produced by van Gompel et al. (2001) for certain NP/VP attachment ambiguities, and by van Gompel et al. (2005) for NPs postmodified by progressive participial VPs. Traxler et al. (1998); van Gompel et al. (2001, 2005) have argued that these reading patterns are problematic for parallel competition-based accounts, according to which an ambiguity left unresolved should, if anything, give rise to greater and longer-lasting difficulty. These reading patterns are, however, predicted by the parallel surprisal-based account, as will now be demonstrated.

For the sentences in (7) it seems uncontroversial to follow Traxler et al. (1998) in assuming that the two main structural alternatives up through the words *... who shot* are essentially as shown in Figure 4, with a partially constructed relative clause attaching either high or low



(a) High-attached relative clause (RC_{high}) (b) Low-attached relative clause (RC_{low})

Figure 4: High versus low attachments of the relative clause in (7)

into a PP-modified NP. This attachment ambiguity is relatively equibiased, so the conditional probability given the string prefix will be substantial for each of these partial structures.[28] The conditional *word* probability of a reflexive pronoun (either *himself* or *herself*) in this context is simply the weighted sum of the probability of the pronoun given each partial structure.

We can write out the equations specifying the conditional probability of *himself* for (7-b) and (7-c) as follows:

$$P_i(\textit{himself}) = P_i(\text{RC}_{\text{low}})P(\textit{himself}|\text{RC}_{\text{low}}) + P_i(\text{RC}_{\text{high}})P(\textit{himself}|\text{RC}_{\text{high}})$$

It seems safe to assume that the probabilities of high versus low RC attachment, $P_i(\text{RC}_{\text{low}})$ and $P_i(\text{RC}_{\text{high}})$, are both substantial (since the attachment is fairly equi-biased), and are approximately equal for (7-b) and (7-c). Likewise, the term $P(\textit{himself}|\text{RC}_{\text{low}})$, which represents the probability that a relative clause modifying the lower NP *the colonel* and beginning with *who shot...* will continue with *himself*, should be approximately equal for the two stimuli. The term $P(\textit{himself}|\text{RC}_{\text{high}})$, however, varies dramatically between (7-b) and (7-c): for the latter, its magnitude should be on the order of $P(\textit{himself}|\text{RC}_{\text{low}})$ (following the reasoning that *the son of the colonel shot himself* and *the colonel shot himself* are similar-probability events), but for the former it is zero, because the word *himself* simply cannot appear in a position where its antecedent must be *daughter*. $P_i(\textit{himself})$ can thus be expressed as wx for (7-b), and $wx + yz$ for (7-c), where the pairs w, y and x, z are of similar magnitude. The conditional probability of *himself* is therefore substantially higher (around double) in (7-c), as both attachments contribute probability mass to the continuation, than in (7-b), where only the low attachment contributes probability mass. Similar reasoning can be applied to (7-a), where *herself* receives probability mass from only the high attachment.

[28] Tree searches in the Brown corpus revealed 9 such examples of sentence-initial high attachments, and 15 of low attachments. WSJ corpus figures were more strongly low-biased at 7 : 28, but included a much larger number of partitives such as *fully 80% of employees* and *nearly all of the crude oil* which, when modified by a non-restrictive relative clause, are obligatorily low attachments. The Brown corpus frequencies comport well with forced-choice offline attachment preferences determined by Traxler et al. (1998), who found a 70% preference for low attachment.

Quite recently, Green and Mitchell (2006) have argued that the results of Traxler et al. (1998); van Gompel et al. (2001, 2005) are actually unproblematic for competition models, on the basis of an extensive set of simulations using the Spivey-Knowlton (1996) normalized recurrence algorithm serving as the basis for prominent competition-based modeling results such as McRae et al. (1998), Spivey and Tanenhaus (1998), and Ferretti and McRae (1999). Green and Mitchell show that the *mean* level of attachment preference across a set of experimental materials is insufficient to determine the predictions of a competition model: the *distribution* of attachment preference within the materials is also important. If the variance in attachment preference is large, then a competition model will essentially mimic a serial model: it will attach low for low-preference items and high for high-preference items, and will thus be garden-pathed some of the time in each disambiguating condition but never in the undisambiguated condition.[29] This defense of competition-based models differs from surprisal's explanation: under surprisal, the undisambiguated condition would remain easier even if all items were perfectly equibiased at all points in the input prior to the critical region.

The parallel surprisal theory's explanation of unresolved-ambiguity data differs markedly from the serial variable-choice model advanced by Traxler et al.. In the variable-choice account, the parser, upon encountering the relative pronoun *who*, stochastically chooses either high or low attachment and continues on with a serial parse, backtracking only if that serial parse subsequently fails (as happens if the low attachment was chosen in (7-a), or the high attachment in (7-b)). This is a claim that the parser is garden-pathed some of the time, and that the observed differential difficulty results from reanalysis during some trials on (7-a) and (7-b). The variable-choice account would seem a natural explanation for the data in Traxler et al. (1998) and van Gompel et al. (2001), where the critical disambiguation site appears several words downstream of the attachment site. Variable choice might be a less plausible explanation for the data in van Gompel et al. (2005), where the site of disambiguation is the first word of the left-attaching phrase, as in (8) below. van Gompel et al. found the undisambiguated condition (8-c) easier than either disambiguated condition, parallel to the results Traxler et al. (1998) found for (7).

(8) a. I read that the governor of the province retiring after the troubles is very rich.
b. I read that the province of the governor retiring after the troubles is very rich.
c. I read that the bodyguard of the governor retiring after the troubles is very rich.

In these stimuli, an attachment decision can only be made after the critical word *retiring* is recognized as initiating a phrase that can be left-attached to an NP. Unlike (7), however,

[29]This analysis would apply equally, of course, to distribution of attachment preference among participants. Green and Mitchell (2006) also make the point that the McRae et al. (1998) version of Spivey-Knowlton (1996)'s model actually starts disambiguating the attachment before the modifier is ever encountered, which, in combination with the model's "rich get richer" dynamical feedback mechanism, effectively magnifies small attachment preference differences. However, the fact that the McRae et al. (1998) model begins disambiguating modifier attachments before encountering the modifiers may perhaps be considered an idiosyncracy that might not apply more broadly to competition models in general.

at the moment when the attachment decision can first be made in (8), the parser has all the information necessary—the head nouns of both attachment sites together with the disambiguating word—always to make the correct attachment decision. For a rational parser not to avoid reanalysis in these situations would require a considerably impoverished parsing regimen where little to no top-down information is available, a requirement which seems to fly in the face of the capacity to integrate a variety of contextual information into attachment decisions (Tanenhaus et al., 1995). The parallel surprisal theory, on the other hand, deals with the data in (8) unproblematically, since the word *retiring* in (8-c) derives probability mass from the possibility of modification of either preceding NP, whereas in (8-a) and (8-b) it derives probability mass from only one possible attachment. The facilitative effect of ambiguity under surprisal also turns out to be important in the analysis of the German subject preference presented in Section 7.

There are several other ways in which the basic manipulation might be varied to tease apart surprisal and variable-choice models. Variable choice, but not surprisal, predicts bimodality in response measures at the critical region (see also Gibson and Pearlmutter 2000). Under variable choice we might also expect that as the amount of material intervening between the ambiguous attachment and the disambiguating region is increased, the observed relative difficulty incurred in ambiguity resolution would increase, if we introduce the assumption that recovery is more difficult for a garden path further pursued.[30] Finally, note that the analysis in this section relied on the assumption that $P(w_i|T)$ was approximately equal for different T. If the experimental contrast is altered so as to break this assumption, we can cause the predictions of surprisal to diverge from those of variable choice. Consider the contrast in (9) below, for example.

(9) a. The suicidal daughter$_{\text{i}}$ of the colonel$_{\text{j}}$ who shot herself$_{*\text{i/j}}$ on the balcony had been very depressed.

b. The homicidal son$_{\text{i}}$ of the colonel$_{\text{j}}$ who shot himself$_{\text{i/j}}$ on the balcony had been very depressed.

In a generative probabilistic model sensitive to the pragmatics and lexical semantics of the words *suicidal* and *homicidal*, we would expect the probability of *herself* given high attachment in (9-a) to greatly exceed the probability of *himself* given either attachment in (9-b).[31] As long as this manipulation of the high NP does not drastically affect the prior probability of high versus low attachment, the conditional word probability of *herself* in (9-a) would then be much higher than that of *himself* in (9-b), but the variable-choice model would in contrast predict that average reflexive pronoun difficulty should still be higher in (9-a), since it is only in this stimulus that the possibility of being garden-pathed exists at all.

[30] Introducing such an assumption might also be a way of handling the length-sensitive "digging-in effects" mentioned in Section 8.3 within the variable choice model.

[31] This hypothesis might be tested via Cloze completions of the partial sentence *The* {*colonel/suicidaal woman/homicidal man*} *shot* __.

7 The subject preference

Variable word order in natural languages can give rise to local ambiguities involving which grammatical function (GF) is assigned to a particular noun phrase. Such local ambiguity is possible in a wide variety of languages: although languages with free word order often use case to mark GFs on noun phrases, syncretism of case form across multiple GFs is also widespread, being documented in Australian, Finno-Ugric, Indo-European, and Turkic languages (e.g., Carstairs 1984; Comrie 1978, 1986; Kiparsky 2001). The situation with respect to online processing is best documented in German (e.g., (Hemforth, 1993; Schlesewsky et al., 2000; Bornkessel et al., 2002)). For example, when a clause-initial NP is syncretized between nominative and accusative case, as in (10) below, there is a temporary ambiguity between subject and object interpretations of that initial NP. The postverbal NP is read more quickly when it is accusative (10-a) than when it is nominative (10-a), indicating a default subject preference (Hemforth, 1993).

(10) a. die Henne sieht den Bussard
the $\text{hen}_{\text{NOM/ACC}}$ sees the_{ACC} buzzard
"The hen sees the buzzard."

b. die Henne sieht der Bussard
the $\text{hen}_{\text{NOM/ACC}}$ sees the_{NOM} buzzard
"The buzzard sees the hen."

This preference is relevant to both frequency- and locality-based parsing theories, because one "default" word order (subject before object for German) is often much more frequent than the other, and movement-based syntactic theories of the alternate orderings can create locality asymmetries. This section closely investigates two experiments conducted by Schlesewsky et al. (2000) which are particularly interesting because they investigate a granularity level at which construction-frequency accounts along the lines of the Tuning Hypothesis may be ruled out. They point out that although declarative clauses in general are usually subject-initial, interrogative clauses beginning with the inanimate, case-syncretized word *was* 'what' may not be. They report that of 480 sentence-initial *was* 'what' items randomly selected from the Freiburg Corpus, about 55% were accusative, suggesting that frequency-based considerations should not favor a default subject interpretation for inanimate case-syncretized initial NPs.

Schlesewsky et al. also consider a movement-based syntactic account along the lines of the Active Filler Hypothesis. In their account, main declarative clause order is derived by movement of the finite verb and an argument NP from an underlying SOV order to the head and specifier positions of CP, respectively. In (10), as soon as the parser has seen the finite verb *sieht*, it can posit an immediately following gap in the subject position and resolve it with the sentence-initial filler, as in Figure 5. Under this preferred parse, however, the next NP cannot be nominatively marked, so (10-b) will cause processing difficulty. Thus the AFH predicts the subject preference independent of construction frequency.

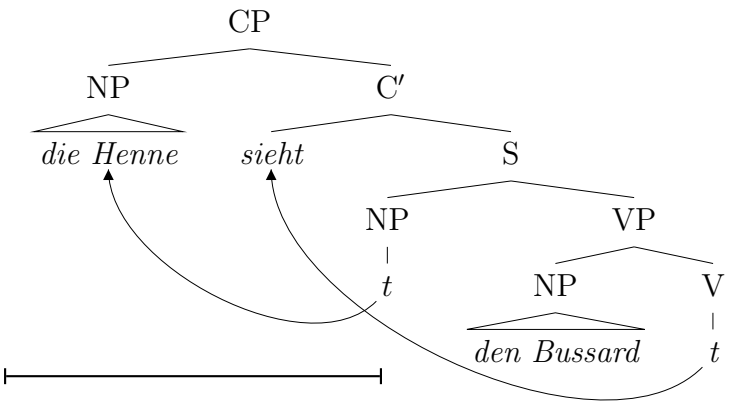


Figure 5: Schlesewsky et al. (2000)'s Subject Preference in German declarative clauses (Example (10-a)), as derived by the AFH and a movement-based analysis of German clause order. For Example (10-b), *Bussard* has the nominative article *der*, and the greedy assignment of *sieht* to the V gap creates a case marking conflict.

Schlesewsky et al. (2000) conducted two experiments involving singular, neuter, case-syncretized sentence-initial *wh* words, to determine whether a subject preference persists when construction-frequency differentials are neutralized. In one experiment, disambiguation involves number marking on the main verb (11); in the other, disambiguation occurs via case marking on the postverbal NP (12).

(11) welches System | **unterstützt/unterstützen** | die Programme | auf den Computer | ?
which system | supports/support | the programs | on the computer | ?
"Which system {supports the programs on the computer/do the programs on the computer support} ?"

(12) was | erforderte | **den/der** Einbruch | in die Nationalbank | ?
what | required | the.ACC/.NOM break-in | into the national_bank | ?
"What {required the break-in into the national bank/did the break-in to the national bank require}?"

In (11), verbal agreement in the plural *unterstützen* condition disambiguates the grammatical function of the sentence-initial singular neuter NP *welches System*; in the singular *unterstützt* condition, disambiguation occurs at the (nom/acc syncretized) postverbal NP, which being plural cannot be the subject of a singular verb. In a serial model, default subject intepretation preference for the initial NP predicts greater processing difficulty at the main verb in the *unterstützen* condition. Schlesewsky et al. (2000) confirmed this prediction experimentally, finding higher reading time for the plural *unterstützen* condition of (11) starting at the main verb and persisting through the rest of the sentence.

	was			*welches* + N		
	Subj	Obj	Other	Subj	Obj	Other
NEGRA	43	18	19	0	0	0
TIGER	84	47	23	0	1	0
TüBa-D/Z (Nom/Acc)	40	37	18	1	0	0

Table 4: Empirical frequencies of subject and object interpretations of sentence-initial *was* and *welches*

In (12), verbal agreement is singular, meaning that the verb is compatible with either a subject or object reading for sentence-initial *was*. Case marking on the immediate postverbal NP is unambiguous, however, and disambiguates the grammatical function of the clause-initial NP *was*. In a serial model, default preference for subject interpretation of the sentence-initial NP would predict greater processing difficulty at the postverbal NP in the *der* condition. Schlesewsky et al. (2000) indeed found significantly higher reading time for the *der* condition, but at the postverbal NP the difference was small and statistically insignificant; it reached significance (as well as its largest numerical difference) at the postmodifying PP.

As discussed in Section 4.4, the predictions of surprisal can differ substantially from construction-frequency accounts such as the Tuning Hypothesis when dependents precede their heads, as is the case for subject-preference data. The remainder of this section presents a surprisal-based analysis of the data in (11) and (12). First, however, it is instructive to use readily-available hand-parsed corpora of German to determine the generality of the corpus-frequency counts of Schlesewsky et al. (2000). Table 4 shows these counts for the NEGRA corpus, as well as for two other parsed corpora of German, TIGER and TüBa-D/Z.[32] We see a considerable corpus-dependent difference: for Frankfurter Rundschau text (NEGRA and TIGER), there is a clear trend toward greater frequency of subject for sentence-initial *was*, but for Die Tagezeitung text (TüBa-D/Z), as with the reported Freiburg Corpus counts, subject and object seem to be similar in probability as GFs for initial *was*.

7.1 *welches* questions with disambiguating agreement

The intuitive difference between surprisal and serial construction-frequency accounts of (11) becomes clear when the full set of structural continuations of *Welches System...* that could lead to the finite verb is examined:

(13) a. [Welches System]$_{\text{SUBJ}}$ V.sg ...

b. [Welches System]$_{\text{OBJ}}$ V.sg ...

[32]TIGER, like NEGRA, is a hand-parsed corpus of text from the German newspaper Frankfurter Rundschau (Brants et al., 2002). TüBa-D/Zis a hand-parsed corpus of text from the German newspaper Die Tagezeitung, which is more colloquially written than Frankfurter Rundschau (Telljohann et al., 2005).

c. [Welches System]$_{\text{OBJ}}$ V.pl ...

d. *[Welches System]$_{\text{SUBJ}}$ V.pl ...

As discussed in Section 6, surprisal *marginalizes* over multiple structural interpretations of a partial input to determine the expectation of the next word. Since a sentence-initial object does not constrain number marking on the upcoming finite verb, singular verb expectations receive probability mass from not only subject ((13-a)) but also object (13-b)) interpretations of the clause-initial NP. Plural verb expectations, in contrast, receive expectation from only the object interpretation of the clause-initial NP (note that continuation (13-d), because it violates subject-verb agreement, will receive little to no expectation from a rational probabilistic model.) Informally, then, even if the probability of an object interpretation of clause-initial *Welches System* is over 0.5, the expectation it contributes to finite verbs is split between singular and plural verbforms. In the case-marked NEGRA corpus, 60.6% of clause-initial objects are in fact followed by a singular finite verb, whereas only 15.2% are followed by a plural finite verb. This causes the surprisal to be greater for plural verbs than for singular. Two detailed quantitative estimates of finite-verb surprisal differences for (11) are given in Appendix B in support of this informal analysis.

7.2 *was* questions with disambiguating case marking

We now examine the stimuli in (12), in which the grammatical function of sentence-initial *was* is disambiguated by case marking on the definite article of the immediately postverbal NP. We therefore will investigate word-by-word surprisal differentials arising from a case-marked PCFG derived from the NEGRA corpus, just as in Section 5.2.[33] Figure 7 plots the word-by-word differences in the subject-*was* and object-*was* conditions for (a) surprisal, based on the case- and number-percolated PCFG read off the morphologically annotated portion of NEGRA; and (b) actual mean reading time of the word's region in (12).[34] The scaling factor of the graph is the slope of a linear regression (with zero intercept) of by-region reading time against surprisal.

The crucial points in Figure 7 are the surprisals and reading times at the postverbal article *der* and at the preposition *in*. Because these are closed-class words occurring in high-frequency syntactic contexts, we can be relatively confident that statistical variance in

[33]Note that case marking is not represented on the PPs in Figure 6 because their case is not fully constrained before the subsequent NP has been seen. In effect, the PP category in these partial parse trees is shorthand for all the possible case-marked PP categories consistent with the head preposition *in*.

[34]Because the words *erforderte*, *Einbruch*, and *Nationalbank* do not appear in the case-marked portion of NEGRA, it was impossible to measure all relevant probabilities associated with these words. Instead, I have substituted the semantically related words *begründete* 'caused', *Eintritt* 'entrance', and *Bank* 'bank' have been respectively substituted. Since both the substituted words and their replacements are part-of-speech unambiguous, the substitutions have no effect on the surprisal differentials at other words in the sentences determined by an unlexicalized PCFG.

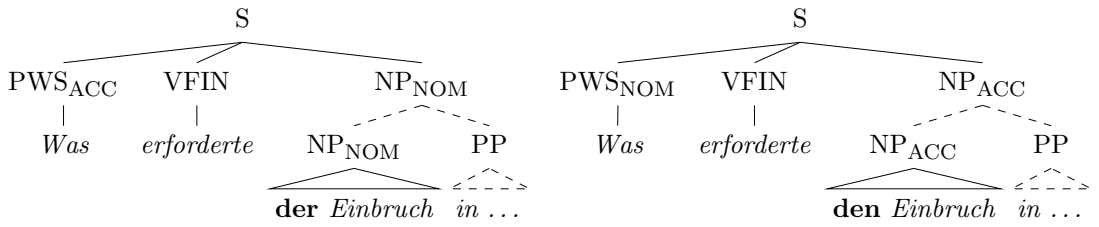


Figure 6: The two major partial parses for (12), with case- and number-percolated categories. (PWS is the part of speech assigned to the word *was* in the NEGRA treebank.)

the surprisal difference across conditions at these points is relatively low.[35] Although the surprisal in the **der** condition of (12) at the postverbal article is indeed higher (by 0.32 bits), it is the postnominal preposition that sees the greatest surprisal differential—over twice as high at 0.70 bits. This pattern matches the empirical results of Schlesewsky et al. (2000), where the greatest difficulty was found at the postnominal PP, not at the postverbal NP.

Two questions now need to be answered regarding the surprisal model's results: why there is a considerable surprisal differential at the postnominal PP, and why there is only a small surprisal differential at the postverbal NP. The first question turns out to have a simple answer upon inspection of incremental parsing under the case-marked PCFG. Figure 6 shows the partial parses leading to the preposition *in* in the **der** and **den** conditions of (12), respectively. The only difference among the grammatical rules required to extend the partial parse through *Einbrugh* to accomodate the new word *in* is the PP adjunction rule:

(14) $\text{NP}_{\text{NOM}} \rightarrow \text{NP}_{\text{NOM}}\ \text{PP}$

versus

(15) $\text{NP}_{\text{ACC}} \rightarrow \text{NP}_{\text{ACC}}\ \text{PP}$

In German, object NPs are empirically more likely than subject NPs to be postmodified by prepositional phrases. This is shown in Table 5: it is true not only of subject versus object NPs overall, but also specifically of subject versus object NPs in the immediate postverbal position.[36] This means that the probability of the rule (14) is higher than the probability of (15). In the online comprehension of (12), immediately after hearing *Einbruch* the comprehender therefore has a greater expectation of seeing a PP (and hence a preposition) next in the *den* condition than in the *der* condition. Hence the surprisal at *in* is greater in the *der* condition.

[35]The surprisal differences at the open-class words *Eintritt* and *Bank*, in contrast, are likely to have high variance, because the prior frequency of the word appearing in different case forms will affect surprisal differentials, and the low count of these words in the case-marked portion of NEGRA (3 instances of *Eintritt*, 7 of *Bank*) causes high variance in the prior-frequency estimate.

[36]All subject/object differentials in Table 5 are significant by Fisher's exact test. For all corpora, postmodification also remains more frequent for object than for subject NPs when only PPs headed by the word *in* are considered, although only the figures for post-verbal NPs (not the figures for all NPs) are statistically significant.

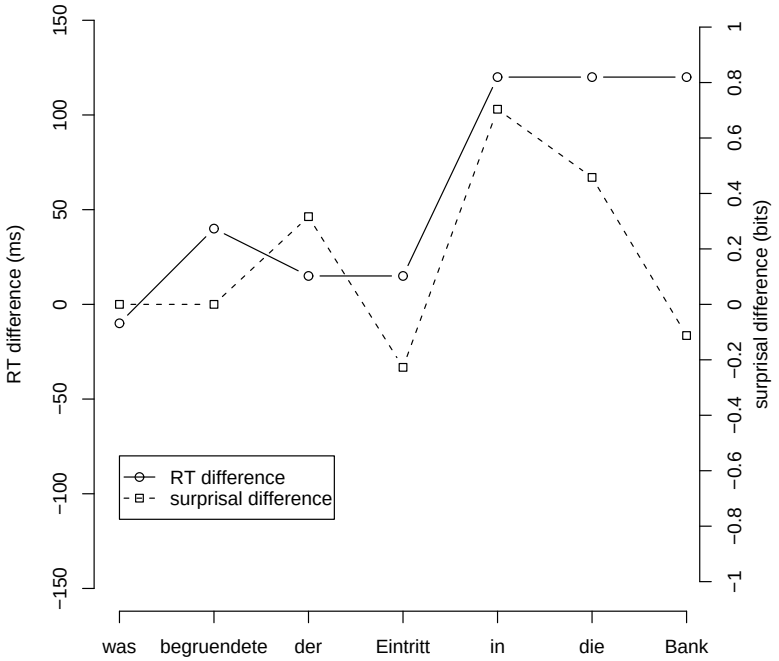


Figure 7: Predicted vs. actual reading time differentials for (12)

The explanation for the small surprisal differential at the onset of the postverbal NP, despite the strong differential frequency of initial-NP grammatical function reported for NEGRA in Table 4, is as follows. First, not all German finite clauses beginning with subject *was* are transitive. Second, in transitive clauses of written German there seems to be an overall tendency to put the subject NP immediately after the finite verb when the object NP is scrambled to initial position: in NEGRA, 73% of subject-initial clauses have an immediately post-verbal object, whereas 90% of object-initial clauses have an immediately post-verbal subject.[37] These two factors conspire to reduce the surprisal advantage of **den** over **der**. Unlike the Active Filler Hypothesis, therefore, surprisal predicts maximal processing difficulty in this experiment precisely where it occurs.

8 Empirical difficulties for the theory

In Sections 5, 6, and 7 we have seen cases where surprisal makes predictions consistent with online processing data that may be difficult to reconcile with other theories. This section

[37]Unfortunately, insufficient data exists to determine whether this pattern extends to *was*-initial clauses in particular.

	All				Postverbal			
	Subj		Obj		Subj		Obj	
	N	%	*N*	%	*N*	%	*N*	%
NEGRA	15220	15.3	7952	22.4	2393	12.2	1156	20.3
TIGER	30187	15.2	17490	23.7	4231	12.2	2461	24.4
TüBa-D/Z	20072	6.3	10094	11.2	5672	4.8	2710	8.6

Table 5: Frequency of PP modification for subject versus object NPs

touches on empirical data that may be difficult for surprisal, and suggests what support for other types of processing theories can be drawn from these data.

8.1 English relative clauses

At this point it is appropriate to return to the configuration that has been most extensively investigated in the context of syntactic processing difficulty: relativization. As noted in Section 4.2, it is well-established that object-extracted RCs in English are more difficult than subject-extracted RCs. In locality-based theories, this is due to the fact that subject but not object relativizations minimize the distance between the extraposition and both the gap and the governing verb. As shown by Hale (2001), surprisal predicts the same general asymmetry due to the fact that object RCs are less common than subject RCs. However, different theories disagree on exactly *where* the increased difficulty of object RCs is predicted to occur, and more recent studies have begun to address this issue by looking at word-by-word reading time patterns in greater detail.

To begin the analysis, note that the integration-cost component of the DLT predicts that it is the RC verb that will be harder to read in object extractions than in subject extractions, because the verb (and the immediately postverbal gap) is where the extra integration cost is paid. Within the surprisal theory, on the other hand, a relative pronoun triggers a syntactic environment much like a verb-final clause: the comprehender knows that the RC's verb must appear at some point, but is uncertain as to what it is and whether a subject will precede it. Surprisal therefore predicts that RC verbs should be read more *slowly* in subject RCs than in object RCs. The cost of low expectation for object RCs should be paid at the embedded subject, which is where the bulk of the expectation devoted to seeing a subject-extracted RC is pruned away.[38] But the empirical evidence in this case seems to side with locality over surprisal. Grodner et al. (2000) show that for stimuli of the form in (16) below, there is a marked increase in the reading time at the embedded verb *sent* for the object over the subject relativization. The embedded subject in (16-b), *the photographer*, is read quickly (see Appendix B of Grodner and Gibson (2005) for word-by-word reading times):

(16) a. The reporter who sent the photographer to the editor hoped for a good story.

[38]The DLT's storage component and the AFH both predict a degree of cost at the embedded subject in an object relativization, but these predictions have no baseline of comparison and at any rate turn out to be inferior in granularity to the predictions of surprisal, so I will not discuss them further.

b. The reporter who the photographer sent to the editor hoped for a good story.

One possible interpretation of this result within the surprisal theory would be that the observed slowdown at the main verb is a *spillover* effect: the difficulty is actually incurred at the embedded subject NP, but it is not registered until the embedded verb. Two natural ways of testing this interpretation present themselves. First, the distance between the embedded subject and the embedded verb could be increased: a spillover effect should occur on the material right after the embedded subject NP, whatever it happens to be. Alternatively, the surprisal theory could be tested for by modulating the the embedded subject NP so that it is more or less predictable. A more predictable embedded subject NP should be read more quickly than one that is less predictable.

An experiment relevant to the spillover prediction was conducted by Grodner and Gibson (2005), who varied postmodification of the subject NP in embedded RC context:

(17) a. The administrator who the nurse supervised...
b. The administrator who the nurse **from the clinic** supervised...
c. The administrator who the nurse **who was from the clinic** supervised...

The DLT predicts that the difficulty of the first verb will be lowest in the unmodified case, higher in the PP-modified case, and highest in the RC-modified case. Surprisal predicts exactly the reverse pattern; and furthermore, if the embedded-verb difficulty seen in (16-b) is due to spillover from the embedded subject NP, we might expect to see a spillover spike inside the postmodifiers of (17-b) and (17-c). The experimental results in this case generally support the DLT: RC-verb reading time is elevated significantly in (17-c), and in (17-b) the PP *from the clinic* is consistently read quickly, which undermines a spillover account of verbal difficulty in (16-b).

Gordon et al. (2004) provide another piece of the puzzle by varying the definiteness and quantification of embedded subject NPs in object-extracted RCs. The crucial contrasts involve the following stimulus types:

(18) a. The salesman that **{the/an} accountant** contacted spoke very quickly. (Definite/Indefinite)
b. The salesman that **(the) accountants** contacted spoke very quickly. (Definite/Bare Plural)
c. The salesman that **{the accountant/everyone}** contacted spoke very quickly. (Definite/Quantifier)

In a corpus study within the same article, the authors found definite NPs to outnumber their indefinite or bare counterparts for both singular and plural embedded subjects. To reason about the predictions of surprisal for these cases, it is necessary to recall that the theory links processing difficulty to the conditional probability of each *word* in its context. This encompasses lexical probabilities, so a rare word as a syntactically likely continuation may well be more surprising than a common word as a syntactically unlikely continuation. In the (18-a) contrast, the discrepancy in definite/indefinite NP frequency is the only relevant

statistic, so surprisal predicts that the definite NPs should be easier. Surprisal also predicts that the definite NPs should be easier in (18-b), but the difference in difficulty should be more dramatic, because the comprehender receives more information at once—both the fact of an object RC and the main lexical content of the embedded subject—in the bare plural case than in the indefinite singular case. In the (18-c) case, the relevant contrast is likely to be between open-class and closed-class (hence high-frequency) lexical NP head, so we predict lower difficulty for the *everyone* stimulus. These predictions are fairly consistent with the experimental results of Gordon et al. (2004): (18-a) produced no significant differences in reading times, (18-b) produced significantly faster reading times at the embedded subject NP for the definite stimulus (and at the matrix verb, though curiously not at the embedded verb), and in (18-c) reading time was significantly lower at the quantifier NP and beyond.

Taken together, recent results on constrained syntactic environments and English relative clauses pose a perplexing set of results. On the one hand, in verb-final and English matrix-verb environments, extra dependencies preceding the head seem to facilitate rather than hinder reading at the final verb, as we saw in Section 5. On the other hand, additional and more informative material before the verb of an object-extracted RC seems to hinder, not facilitate, reading time at that verb. Nevertheless, subregularities in the difficulty of embedded subject NPs observed in Gordon et al. (2004) are consistent with the predictions of surprisal.

One way of interpreting these mixed results is to hypothesize that surprisal has a major effect on word-by-word processing difficulty, but that truly non-local (i.e., long-distance) syntactic dependencies such as relativization and *wh*-question formation are handled fundamentally differently from local syntactic dependencies, and the retrieval and integration of a long-distance dependent incurs a substantial processing cost comparable to the cost of a highly surprising word. On this theory, surprisal effects dominate the processing of verb-final clauses because none of the dependencies are long-distance, but processing a relative clause involves storing, retrieving, and integrating a long-distance dependent, so that relative clause reading times also exhibit substantial DLT-like effects that are not predicted by surprisal. Working out such a two-factor theory would be a non-trivial undertaking beyond the scope of this work, but the most recent available data suggests that formulating and testing such an approach could well be a promising direction for future research on syntactic processing difficulty.

8.2 Digging-in effects

One property of some competition and dynamical models proposed in the literature (see Section 4.3) is that they predict *digging in*: while multiple analyses are posssible, the favored analysis tends to become stronger even in the absence of evidence bearing on the ambiguity. One type of evidence that seems to support this idea is a finding by Ferreira and Henderson (1991) recently elucidated and modeled by Tabor and Hutchins (2004) that the difficulty in recovery from a so-called "NP/Z" ambiguity (as in (19) below) increases with the length of the ambiguously-attached NP

(19) a. As the author wrote the book grew.
b. As the author wrote the book describing Babylon grew.

Ferreira and Henderson found (using relative clause rather than gerund VP postmodifiers) that participants judged NP/Z sentences grammatical less often when the ambiguous NP was long, as in (19-b), than when it was short, as in (19-a). Tabor and Hutchins replicated this finding, and in a self-paced reading study showed that the longer NP induced considerably increased processing difficulty at the disambiguating word *grew*. Tabor and Hutchins interpreted this finding as a "digging-in" effect: in the absence of additional information contributing to ambiguity resolution, initial attachment preferences get stronger and stronger, so that the dispreferred subject interpretation of the ambiguous NP becomes increasingly less accessible as the NP increases in length.

As presented here, surprisal does not predict digging-in effects: there is no time-dependent positive-feedback process invoked during incremental sentence comprehension. However, this does not mean that the NP/Z results described above are incompatible with the theory. The reason for this is that the size and structure of the ambiguous NP *does* constitute potentially disambiguating information. In English, object NPs are typically larger than subject NPs. In the parsed Brown corpus, for example, subject NPs contain an average of 1.87 words, object NPs an average of 4.20 words (if pronominal NPs are excluded, the figures are 2.77 and 4.95 respectively). On the basis of arguments made in Sections 3 and 5.1, it is reasonable to expect this information to be deployed in incremental disambiguation. As a result, the postmodifier in (19-b) should strengthen the preference for object interpretation of the ambiguous NP, and correspondingly increase predicted difficulty of the disambiguating verb "grew".

8.3 Local coherence effects

Another type of result that has received considerable recent attention and is what could be called *local coherence effects*: when difficulty arises from a source that seems to be independent of or even violate constraints imposed by possible structures or structural preferences imposed by the global (i.e., complete incremental sentence) context. Tabor et al. found that when a reduced relative clause modifying a noun within an unambiguously non-subject context is introduced by a verb that is part-of-speech-ambiguous between past participle and simple past (such as *tossed*), additional processing difficulty relative to that incurred for an unambiguously simple-past verb (such as *thrown*) is incurred:

(20) a. The coach smiled at the player thrown the frisbee.
b. The coach smiled at the player tossed the frisbee.

Within a fully incremental probabilistic theory of comprehension, it would be possible to entertain these locally-coherent analyses (i.e., *the player tossed* as a subject-verb combination) by loosening the set of constraints on what constitutes a well-formed tree, leaving a small amount of probability mass for "marginal" analyses that are not completely in concord with all categorical grammatical constraints. This step would not allow surprisal to explain the observed result, however, because these marginal analyses would, if anything, contribute

more probability mass to *tossed* than to *thrown*. Hence, example (20) would become a case of facilitative ambiguity, as in Section 6, and the *tossed* condition should be easier than the *thrown* condition. Yet the opposite is observed.

This piece of data is therefore a point of empirical difficulty for the surprisal theory as presented here. One possible starting point for an analysis, however, could be to explicitly introduce uncertainty about *previous words in the sentence* into the model. In any instance of sentence comprehension, the previous words in the sentence must be retained in short-term memory, and the comprehender must retain a degree of uncertainty as to exactly what those words were.[39] It is notable that small edits in the structure of (20-b) make the locally-coherent reading globally coherent as well:

(21) a. The coach smiled at ***how*** the player tossed the frisbee.
b. The coach smiled at the player ***who/that*** tossed the frisbee.

On the reranking interpretation of surprisal, the difficulty of the critical word *tossed* could be due to reranking it induces on the distribution over previous words in the sentence. The critical word *thrown* in (20-a) would not induce a corresponding reranking, because it is incompatible with the edits in (21). Elucidating such an analysis is, however, beyond the scope of the present paper.

9 Conclusion

Recent experimental results in syntactic ambiguity resolution indicate that comprehenders incrementally integrate a variety of evidential knowledge in the process of discriminating the preferred interpretation of a sentence; probability theory serves as a coherent architecture for this *constraint-based*, *resource-allocation* paradigm of ambiguity resolution. We can extend the parallel, probabilistic disambiguation perspective of incremental sentence processing into a theory of syntactic complexity and processing difficulty by formalizing a linking hypothesis stating that the primary source of difficulty incurred in processing a given word is determined by the degree of update in the preference distribution over interpretations of the sentence that the word requires. Formalized appropriately using the information-theoretic measure of the *relative entropy* between probability distributions, we are able to derive a theory of processing difficulty previously proposed by Hale (2001), that the difficulty of a word is the *surprisal* (negative log of the conditional probability) of that word given its context. This surprisal theory has several desirable theoretical and mathematical properties, including a coherent integration of rational disambiguation, incremental processing, and differential processing difficulty; its ability to serve as a causal bottleneck between representations and predictions about processing difficulty; and freedom from the granularity bias of other probabilistic theories of syntactic comprehension. Empirically, it can smoothly incorporate major results in the literature involving prediction and ambiguity resolution; it also makes non-

[39] This is perhaps particularly true in non-cumulative presentations, which were used in Tabor et al. (2004) and Konieczny (2005).

trivial predictions about (1) processing difficulty in head-final and similar contexts where the comprehender knows that a certain type of constituent is upcoming, but is uncertain as to exactly *where* and *what* it is; and (2) circumstances under which unresolved ambiguity can facilitate comprehension. As seen in Sections 5 through 8.1, these predictions are for the most part confirmed by existing experimental results.

There are three more general conclusions that we can also draw from the investigation of expectation-based processing theories presented here. One is the utility of causal bottlenecks in theories of syntactic comprehension. At first glance it would seem impossible to talk about syntactic comprehension without making firm commitment to specific syntactic structures; and in fact there has been a history of differentiating predictions about behavioral metrics on the basis of alternative structural representations, from Minimal Attachment (Frazier and Fodor, 1978) to the more recent Entropy Reduction Hypothesis (Hale, 2006). In the surprisal theory, in contrast, structural representations affect processing difficulty only through the mediation of probabilistic word models. The latter can be investigated through a variety of means, potentially including completion as well as comprehension studies. Yet surprisal is not a repudiation of syntax: as we have seen, probabilistic word models can be estimated from probabilistic grammars, and even simple grammars can determine models with difficulty patterns strikingly similar to established experimental results. Furthermore, surprisal does not foreclose the possibility of using psycholinguistic data to help characterize the formal nature of probabilistic grammatical knowledge, as different classes of probabilistic string languages require different formal means of finite expression. In this respect, investigations under the surprisal theory can be thought of as a psycholinguistic analogue to empirical and mathematical investigations into the weak generative capacity of the language faculty in the 1980s (Culy, 1985; Shieber, 1985, inter alia).

In addition, research on syntactic processing and on predictability should keep closer abreast of one another. This conclusion is a direct consequence of the log-scale of the surprisal theory. Most work on prediction has focused on highly predictable words—Cloze probabilities of 0.3 and above. But if the correct scale of predictability effects is logarithmic, then difficulty asymmetries can arise even for words whose Cloze probabilities would require enormous studies to accurately estimate. The surprisal theory in fact relies on difficulty asymmetries between low-probability words to explain results discussed in Sections 5 and 5.3.

Finally, results discussed in Section 8.1 and 8.3 suggest that no one source of processing difficulty can explain all the prominent results in syntactic comprehension. In particular, difficulty asymmetries involving relative clauses seem to support a combination of locality- as well as expectation-based difficulty. Integrating DLT-style locality into a fully-parallel processing theory such as surprisal is, however, far from a trivial task, and would be facilitated by more comprehensive experimental investigation of the circumstances under which evidence exists for both effects together. Out of necessity, future work will focus on the crossroads between these two very different views of how difficulty in sentence comprehension arises.

Acknowledgments

This work has benefited from presentation and discussion at a variety of venues, including the 2005 annual meeting of the Linguistic Society of America and the 18th annual CUNY Sentence Processing Conference. I am grateful to feedback on the manuscript from John Hale, Florian Jaeger, Dan Jurafsky, Andrew Kehler, Frank Keller, Christopher Manning, Don Mitchell, Martin Pickering, Ivan Sag, Tom Wasow, and three anonymous reviewers. I accept full responsibility for all errors and omissions. Part of the research reported in this article was supported by an ESRC postdoctoral fellowship, award reference number PTA-026-27-0944.

References

Altmann, G. T. and Kamide, Y. (1999). Incremental interpretation at verbs: restricting the domain of subsequent reference. *Cognition*, 73(3):247–264.

Anderson, J., Bothell, D., Byrne, M., Douglass, S., Lebiere, C., and Qin, Y. (2004). An integrated theory of the mind. *Psychological Review*, 111(4):1036–1060.

Anderson, J. R. (1990). *The Adaptive Character of Human Thought*. Lawrence Erlbaum.

Bod, R. (1992). A computational model of language performance: Data oriented parsing. In *Proceedings of COLING*.

Booth, T. L. (1969). Probabilistic representation of formal languages. In *IEEE Conference Record of the 1969 Tenth Annual Symposium on Switching and Automata Theory*, pages 74–81.

Bornkessel, I., Schlesewsky, M., and Friederici, A. D. (2002). Grammar overrides frequency: evidence from the online processing of flexible word order. *Cognition*, 85:B21–B30.

Brants, S., Dipper, S., Hansen, S., Lezius, W., and Smith, G. (2002). The TIGER treebank. In *Proceedings of the Workshop on Treebanks and Linguistic Theories*.

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

Charniak, E. (1997). Statistical parsing with a context-free grammar and word statistics. In *Proceedings of AAAI*, pages 598–603.

Charniak, E. (2001). Immediate-head parsing for language models. In *Proceedings of ACL*.

Christiansen, M. H. and Chater, N. (1999). Toward a connectionist model of recursion in human linguistic performance. *Cognitive Science*, 23(2):157–205.

Clifton, C. and Frazier, L. (1989). Comprehending sentences with long distance dependencies. In Carlson, G. and Tanenhaus, M., editors, *Linguistic Structure in Language Processing*, pages 273–317. Dordrecht: Kluwer.

Collins, M. (1999). *Head-Driven Statistical Models for Natural Language Parsing*. PhD thesis, University of Pennsylvania.

Collins, M., Hajic, J., Ramshaw, L., and Tillmann, C. (1999). A statistical parser for Czech. In *Proceedings of ACL*.

Comrie, B. (1978). Definite direct objects and referent identification. *Pragmatics-Microfiche*, 3(1):D3.

Comrie, B. (1986). On delimiting cases. In Brecht, R. D. and Levine, J. S., editors, *Case in Slavic*, pages 86–106. Columbus, Ohio: Slavica.

Crocker, M. and Brants, T. (2000). Wide-coverage probabilistic sentence processing. *Journal of Psycholinguistic Research*, 29(6):647–669.

Cuetos, F., Mitchell, D. C., and Corley, M. (1996). Parsing in different languages. In *Language Processing in Spanish*, pages 145–187. Hillsdale, NJ: Erlbaum.

Culy, C. (1985). The complexity of the vocabulary of Bambara. *Linguistics and Philosophy*, 8:345–351.

Dubey, A. and Keller, F. (2003). Parsing German with sister-head dependencies. In *Proceedings of ACL*.

Ehrlich, S. F. and Rayner, K. (1981). Contextual effects on word perception and eye movements during reading. *Journal of Verbal Learning and Verbal Behavior*, 20:641–655.

Elman, J. (1990). Finding structure in time. *Cognitive Science*, 14:179–211.

Elman, J. L. (1991). Distributed representations, simple recurrent networks, and grammatical structure. *Machine Learning*, 7(2-3):195–225.

Engbert, R., Nuthmann, A., Richter, E. M., and Kliegl, R. (2005). SWIFT: a dynamical model of saccade generation during reading. *Psychological Review*, 112(4):777–813.

Federmeier, K. D. and Kutas, M. (1999). A rose by any other name: Long-term memory structure and sentence processing. *Journal of Memory and Language*, 41:469–495.

Ferreira, F. and Henderson, J. M. (1991). Recovery from misanalyses of garden-path sentences. *Journal of Memory and Language*, 31:725–745.

Ferretti, T. R. and McRae, K. (1999). Modeling the role of plausibility and verb-bias in the direct object/sentence complement ambiguity. In *Proceedings of CogSci*.

Frazier, L. (1979). *On Comprehending Sentences: Syntactic Parsing Strategies*. PhD thesis, University of Massachusetts.

Frazier, L. and Fodor, J. D. (1978). The sausage machine: A new two-stage parsing model. *Cognition*, 6(4):291–325.

Frisson, S., Rayner, K., and Pickering, M. (2005). Effects of contextual predictability and transitional probability on eye movements during reading. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 31(5):862–877.

Gazdar, G., Klein, E., Pullum, G., and Sag, I. (1985). *Generalized Phrase Structure Grammar*. Harvard.

Gibson, E. (1991). *A computational theory of human linguistic processing: memory limitations and processing breakdown*. PhD thesis, Carnegie Mellon.

Gibson, E. (1998). Linguistic complexity: Locality of syntactic dependencies. *Cognition*, 68:1–76.

Gibson, E. (2000). The dependency locality theory: A distance-based theory of linguistic complexity. In Marantz, A., Miyashita, Y., and O'Neil, W., editors, *Image, Language, Brain*, pages 95–126. MIT Press.

Gibson, E., Desmet, T., Grodner, D., Watson, D., and Ko, K. (2005a). Reading relative clauses in English. *Language and Cognitive Processes*, 16(2):313–353.

Gibson, E., Nakatani, K., and Chen, E. (2005b). Distinguishing theories of syntactic storage cost in sentence comprehension: Evidence from Japanese. To appear.

Gibson, E. and Pearlmutter, N. J. (2000). Distinguishing serial and parallel parsing. *Journal of Psycholinguistic Research*, 29(2):231–240.

Gordon, P. C., Hendrick, R., and Johnson, M. (2004). Effects of noun phrase type on sentence complexity. *Journal of Memory and Language*, 51(1):97–114.

Green, M. J. and Mitchell, D. C. (2006). Absence of real evidence against competition during syntactic ambiguity resolution. *Journal of Memory and Language*, 55(1):1–17.

Grodner, D. and Gibson, E. (2005). Some consequences of the serial nature of linguistic input. *Cognitive Science*, 29(2):261–290.

Grodner, D., Watson, D., and Gibson, E. (2000). Locality effects on sentence processing. Presented at the 2003 annual CUNY Sentence Processing Conerence.

Hale, J. (2001). A probabilistic Earley parser as a psycholinguistic model. In *Proceedings of NAACL*, volume 2, pages 159–166.

Hale, J. (2003a). *Grammar, Uncertainty and Sentence Processing*. PhD thesis, John Hopkins University.

Hale, J. (2003b). The information conveyed by words in sentences. *Journal of Psycholinguistic Research*, 32(2):101–123.

Hale, J. (2006). Uncertainty about the rest of the sentence. *Cognitive Science*, 30(4):609–642.

Hemforth, B. (1993). *Kognitives Parsing: Repräsentation und Verarbeitung sprachlichen Wissens*. Sankt Augustin: Infix.

Henderson, J. (2004). Lookahead in deterministic left-corner parsing. In *Proceedings of the Workshop on Incremental Parsing: Bringing Engineering and Cognition Together*.

Itti, L. and Baldi, P. (2005). Bayesian surprise attracts human attention. In *Advances in Neural Information Processing Systems*.

Jaeger, F., Fedorenko, E., and Gibson, E. (2005). Dissociation between production and comprehension complexity. Poster Presentation at the 18th CUNY Sentence Processing Conference, University of Arizona.

Jelinek, F. and Lafferty, J. D. (1991). Computation of the probability of initial substring generation by stochastic context free grammars. *Computational Linguistics*, 17(3):315–323.

Johnson, M. (1998). PCFG models of linguistic tree representations. *Computational Linguistics*, 24(4):613–632.

Jurafsky, D. (1996). A probabilistic model of lexical and syntactic access and disambiguation. *Cognitive Science*, 20(2):137–194.

Jurafsky, D. (2003). Probabilistic modeling in psycholinguistics: Linguistic comprehension and production. In Bod, R., Hay, J., and Jannedy, S., editors, *Probabilistic Linguistics*. MIT Press.

Kaiser, E. and Trueswell, J. C. (2004). The role of discourse context in the processing of a flexible word-order language. *Cognition*, 94:113–147.

Kay, M. (1980). Algorithm schemata and data structures in syntactic parsing. In *Proceedings of the Nobel Symposium on Text Processing*. Gothenburg.

Keller, F. and Lapata, M. (2003). Using the web to obtain frequencies for unseen bigrams. *Computational Linguistics*, 29:459–484.

King, J. and Just, M. A. (1991). Individual differences in syntactic processing: The role of working memory. *Journal of Memory and Language*, 30(5):580–602.

Kiparsky, P. (2001). Structural case in Finnish. *Lingua*, 111(4–7):315–376.

Klein, D. and Manning, C. D. (2003). Accurate unlexicalized parsing. In *Proceedings of ACL*.

Konieczny, L. (1996). *Human sentence processing: a semantics-oriented parsing approach*. PhD thesis, Universität Freiburg.

Konieczny, L. (2000). Locality and parsing complexity. *Journal of Psycholinguistic Research*, 29(6):627–645.

Konieczny, L. (2005). The psychological reality of local coherences in sentence processing. In *Proceedings of the 27th Annual Conference of the Cognitive Science Society*.

Konieczny, L. and Döring, P. (2003). Anticipation of clause-final heads: Evidence from eye-tracking and SRNs. In *Proceedings of ICCS/ASCS*.

Kutas, M. and Hillyard, S. A. (1980). Reading senseless sentences: Brain potentials reflect semantic incongruity. *Science*, 207(4427):203–205.

Kutas, M. and Hillyard, S. A. (1984). Brain potentials during reading reflect word expectancy and semantic association. *Nature*, 307:161–163.

MacDonald, M. C. (1993). The interaction of lexical and syntactic ambiguity. *Journal of Memory and Language*, 32:692–715.

MacDonald, M. C., Pearlmuttter, N. J., and Seidenberg, M. S. (1994). Lexical nature of syntactic ambiguity resolution. *Psychological Review*, 101(4):676–703.

Magerman, D. M. (1994). *Natural Language Parsing as Statistical Pattern Recognition*. PhD thesis, Stanford.

Marcus, M. P., Santorini, B., and Marcinkiewicz, M. A. (1994). Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330.

Marslen-Wilson, W. (1975). Sentence perception as an interactive parallel process. *Science*, 189(4198):226–228.

McDonald, S. A. and Shillcock, R. C. (2003a). Eye movements reveal the on-line computation of lexical probabilities during reading. *Psychological Science*, 14(6):648–652.

McDonald, S. A. and Shillcock, R. C. (2003b). Low-level predictive inference in reading: The influence of transitional probabilities on eye movements. *Vision Research*, 43:1735–1751.

McRae, K., Spivey-Knowlton, M. J., and Tanenhaus, M. K. (1998). Modeling the influence of thematic fit (and other constraints) in on-line sentence comprehension. *Journal of Memory and Language*, 38(3):283–312.

Mitchell, D. C. (1994). Sentence parsing. In Gernsbacher, M., editor, *Handbook of Psycholinguistics*. Academic Press.

Mitchell, D. C., Cuetos, F., Corley, M., and Brysbaert, M. (1995). Exposure-based models of human parsing: Evidence for the use of coarse-grained (nonlexical) statistical records. *Journal of Psycholinguistic Research*, 24:469–488.

Nakatani, K. and Gibson, E. (2003). An on-line study of Japanese nesting complexity. Presented at the 2003 annual CUNY Sentence Processing Conerence.

Narayanan, S. and Jurafsky, D. (1998). Bayesian models of human sentence processing. In *Proceedings of the Twelfth Annual Meeting of the Cognitive Science Society*.

Narayanan, S. and Jurafsky, D. (2002). A Bayesian model predicts human parse preference and reading time in sentence processing. In *Advances in Neural Information Processing Systems*, volume 14, pages 59–65.

Rayner, K., Ashby, J., Pollatsek, A., and Reichle, E. D. (2004). The effects of frequency and predictability on eye fixations in reading: Implications for the E-Z Reader model. *Journal of Experimental Psychology: Human Perception and Performance*, 30(4):720–732.

Rayner, K. and Well, A. D. (1996). Effects of contextual constraint on eye movements in reading: A further examination. *Psychonomic Bulletin & Review*, 3(4):504–509.

Reichle, E. D., Pollatsek, A., Fisher, D. L., and Rayner, K. (1998). Toward a model of eye movement control in reading. *Psychological Review*, 105(1):125–157.

Roark, B. (2001). Probabilistic top-down parsing and language modeling. *Computational Linguistics*, 27(2):249–276.

Rohde, D. (2002). *A Connectionist Model of Sentence Comprehension and Production*. PhD thesis, Carnegie Mellon University.

Rumelhart, D. E., Durbin, R., Golden, R., and Chauvin, Y. (1995). Backpropagation: The basic theory. In Chauvin, Y. and Rumelhart, D. E., editors, *Backpropagation: Theory, Architectures, and Applications*, pages 1–34. Lawrence Earlbaum.

Schlesewsky, M., Fanselow, G., Kliegl, R., and Krems, J. (2000). The subject preference in the processing of locally ambiguous WH-questions in German. In Hemforth, B. and Konieczny, L., editors, *German Sentence Processing*. Kluwer.

Shieber, S. M. (1985). Evidence against the context-freeness of natural language. *Linguistics and Philosophy*, 8:333–343.

Skut, W., Brants, T., Krenn, B., and Uszkoreit, H. (1997). Annotating unrestricted German text. In *Fachtagung der Sektion Computerlinguistik der Deutschen Gesellschaft für Sprachwissenschaft*, Heidelberg, Germany.

Spivey, M. J. and Tanenhaus, M. K. (1998). Syntactic ambiguity resolution in discourse: Modeling the effects of referential content and lexical frequency. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 24(6):1521–1543.

Spivey-Knowlton, M. (1996). *Integration of Lingusitic and Visual Information: Human Data and Model Simulations*. PhD thesis, University of Rochester.

Stolcke, A. (1995). An efficient probabilistic context-free parsing algorithm that computes prefix probabilities. *Computational Linguistics*, 21(2):165–201.

Tabor, W., Galantucci, B., and Richardson, D. (2004). Effects of merely local syntactic coherence on sentence processing. *Journal of Memory and Language*, 50(4):355–370.

Tabor, W. and Hutchins, S. (2004). Evidence for self-organized sentence processing: Digging in effects. *Journal of Experimental Psychology: Learning, Memory, and Cognition,*, 30(2):431–450.

Tabor, W., Juliano, C., and Tanenhaus, M. K. (1997). Parsing in a dynamical system: An attractor-based account of the interaction of lexical and structural constraints in sentence processing. *Language and Cognitive Processes*, 12(2/3):211–271.

Tabor, W. and Tanenhaus, M. K. (1999). Dynamical models of sentence processing. *Cognitive Science*, 23(4):491–515.

Tanenhaus, M. K., Spivey-Knowlton, M. J., Eberhard, K., and Sedivy, J. C. (1995). Integration of visual and linguistic information in spoken language comprehension. *Science*, 268:1632–1634.

Taylor, W. L. (1953). A new tool for measuring readability. *Journalism Quarterly*, 30:415.

Telljohann, H., Hinrichs, E. W., Kübler, S., and Zinsmeister, H. (2005). *Stylebook for the Tübingen Treebank of Written German*. University of Tübingen, Seminar für Sprachwissenschaft.

Traxler, M. J., Pickering, M. J., and Clifton, C. (1998). Adjunct attachment is not a form of lexical ambiguity resolution. *Journal of Memory and Language*, 39:558–592.

Uszkoreit, H., Brants, T., Duchier, D., Krenn, B., Konieczny, L., Oepen, S., and Skut, W. (1998). Studien zur performzorientieren Linguistik: Aspekte der Relativsatzextraposition im Deutschen. *Kognitionswissenschaft*, 7:129–133.

van Gompel, R. P. G., Pickering, M. J., Pearson, J., and Liversedge, S. P. (2005). Evidence against competition during syntactic ambiguity resolution. *Journal of Memory and Language*, 52:284–307.

van Gompel, R. P. G., Pickering, M. J., and Traxler, M. J. (2001). Reanalysis in sentence processing: Evidence against current constraint-based and two-stage models. *Journal of Memory and Language*, 45:225–258.

Vasishth, S. (2002). *Working memory in sentence comprehension: Processing Hindi center embeddings*. PhD thesis, Ohio State University.

Vasishth, S. (2003). Quantifying processing difficulty in human sentence parsing: The role of decay, activation, and similarity-based interference. In *Proceedings of EuroCogSci*.

Vasishth, S. and Lewis, R. L. (2006). Argument-head distance and processing complexity: Explaining both locality and anti-locality effects. *Language*, 82(4):767–794.

Wasow, T. (2002). *Postverbal Behavior*. CSLI.

A Definition and estimation of probabilistic context-free grammars

A *probabilistic context-free grammar* (PCFG; Booth 1969) consists of a set of context-free rule rewrites, each of which is associated with a probability between zero and 1. The probability of a given rule $A \rightarrow \alpha$ can be identified with the conditional probability of the rule's right-hand side, α, given the left-hand side A—that is, $P(\alpha|A)$. The probability of a context-free tree in a given PCFG is simply the product of probabilities of all the rules that make up the tree. The probability of a string $w_{1\cdots n}$ is the sum of the probabilities of all the trees whose yield is $w_{1\cdots n}$. The probability of a string prefix $w_{1\cdots i}$ is the sum of the probabilities of all strings that begin with $w_{1\cdots i}$, or equivalently, the sum of probabilities of all trees whose yield begins with $w_{1\cdots i}$. String prefix probabilities can be calculated efficiently as a byproduct of bottom-up or left-to-right parsing algorithms, as specified by Jelinek and Lafferty (1991) or Stolcke (1995).

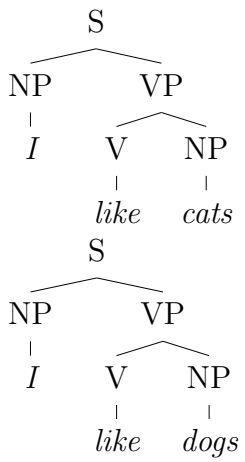
PCFG *estimation* is the process of selecting a set of rules and associated probabilities. A simple form of relative-frequency estimation is employed for all PCFGs used in this paper. Given a collection of syntactic trees (e.g., the Penn or NEGRA Treebank), the number of occurrences of each rule in the collection is counted. The relative-frequency estimate of a given rule R is simply the count of R divided by the total count of all rules whose left-hand side is that of R's:

$$P(A \rightarrow \alpha) = \frac{\text{Count}(A \rightarrow \alpha)}{\sum_{\beta} \text{Count}(A \rightarrow \beta)}$$

Figure 8 gives an example of relative-frequency estimation of a PCFG from a collection of two trees. In the estimated PCFG, the novel tree in the right-hand side of the figure has probability corresponding to the product of the rewrite rules that determine it: $1 \times 1 \times 0.25 \times 0.25 = 0.0625$.

B Analysis of *welches* questions with disambiguating agreement

In the stimuli in (11), the contrasting word of interest is an open-class item whose surface forms, unterstützt and unterstützen, are sparse. In addition, the head noun of the the



Rule	Count	Probability
$S \rightarrow NP\ VP$	2	1
$VP \rightarrow V\ NP$	2	1
$NP \rightarrow I$	2	0.5
$NP \rightarrow cats$	1	0.25
$NP \rightarrow dogs$	1	0.25

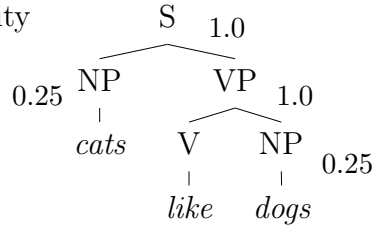


Figure 8: Simple relative-frequency estimation of a PCFG

sentence-initial NP is an open-class word whose relative frequency of occurrence in nominative and accusative forms has a strong effect on the predictions of surprisal predicted by a PCFG. These words are sparse and hence surprisals estimated from corpus-based PCFGs are unlikely to be reliable.

Most straightforwardly, because the region of interest is so close to the beginning of the sentence, we can use the n-gram frequency of the first three words of the sentence to estimate the surprisal at the finite verb directly, given a large corpus of German. A useful estimate of this sort turned out to require using the World Wide Web itself.[40] An exact-match search using Google returned 15 valid matches of the trigram *welches System unterstützt*, many of which were sentence-initial; no instances of the trigram *welches System unterstützen* were found.[41] The direct counting estimate of surprisal at the finite verb therefore predicts the reading-time difference experimentally observed by Schlesewsky et al. (2000).

Alternatively, we can decompose the plausible sources of expectation for the relevant finite verb forms, using grammatical theory and corpus-derived morphosyntactic frequencies. The probability of the singular and plural verb forms *unterstützt* and *unterstützen* in (11) can be decomposed as follows (WS standing for *Welches System*):[42]

$$\begin{aligned} P(\text{unterstützt}|\text{WS}) &= P(\text{V.3sg}|\text{WS})P(\text{unterstützt}|\text{V.3sg}, \text{WS}) \\ P(\text{unterstützen}|\text{WS}) &= P(\text{V.3pl}|\text{WS})P(\text{unterstützen}|\text{V.3pl}, \text{WS}) \end{aligned}$$

[40] See Keller and Lapata (2003) for discussion of issues involved in obtaining n-gram frequencies from the Web.

[41] June 28, 2005, 12:07pm. I discarded one instance of the former trigram that appeared in a web page referencing Schlesewsky et al. (2000).

[42] Although the finite verb forms *unterstützt* and *unterstützen* are compatible with first- and second-person agreement as well, I attend only to third-person agreement because available syntactically-annotated corpora have nearly exclusively third-person subjects. In a model whose parameters more closely reflected speech or another written genre, we might expect $P(\text{V.2}|WS)$ and $P(\text{V.1pl})$, which respectively contribute to the probabilities of *unterstützt* and *unterstützen*, to be substantial.

where V.3sg and V.3pl respectively denote singular and plural third-person finite verbs. This decomposition simply states that the probability of a particular verb-form v given the initial sequence *Welches System* is equal to the probability of a finite verb of the correct number and person marking given the initial sequence, times the probability that the finite verb is actually v. We can make a crude estimate of the second term in the decomposition with the simplifying assumption that the conditioning on *Welches System* does not affect the verb's identity.[43] Under this assumption, the ratio of the right-hand half of the decomposition for *unterstützt* versus *unterstützen* turns out to be roughly 1 : 2.3—the forms themselves are roughly equal in frequency (5 versus 6 in NEGRA, 14 versus 11 in TIGER), and singular finite verbs are roughly 2.3 times as common as plural finite verbs (1844 to 789 in the morphologically-annotated part of NEGRA).

The first term in the decomposition can be further subdivided:

$$\begin{aligned} P(\text{V.3sg}|\text{WS}) &= P(\text{SUBJ}|\text{WS})P(\text{V.3sg}|\text{WS},\text{SUBJ}) \\ &\quad +P(\text{OBJ}|\text{WS})P(\text{V.3sg}|\text{WS},\text{OBJ}) \\ P(\text{V.3pl}|\text{WS}) &= P(\text{OBJ}|\text{WS})P(\text{V.3pl}|\text{WS},\text{OBJ}) \end{aligned}$$

where SUBJ and OBJ refer to the event of sentence-initial NP turning out to be the subject or respectively object of the matrix clause. Crucially, the probability of a singular finite verb has two terms summed together on the right-hand side, because the comprehender can derive expectation for finite verbs from both the subject and object interpretations of the initial NP. The probability of a plural finite verb at this point, on the other hand, has only one term, because a plural finite verb *requires* an object interpretation of the initial NP (so $P(\text{V.3pl}|\text{SUBJ}) = 0$); see also (13). The Freiburg corpus estimates reported by Schlesewsky et al. (2000) for the probablities $P(\text{SUBJ}|\text{WS})$ and $P(\text{OBJ}|\text{WS})$ are 0.45 and 0.55. For the conditional probabilities of V.3sg and V.3pl, make the simplifying assumption of independence between the lexical content of the initial NP and the category of the subsequent constituent:

$$P(\text{V.3}\{\text{sg/pl}\}|\text{WS},\{\text{SUBJ/OBJ}\}) \approx P(\text{V.3}\{\text{sg/pl}\}|\{\text{SUBJ/OBJ}\})$$

These simplified probabilities can be estimated directly from structural counts in the morphologically annotated NEGRA corpus, giving the following estimated probabilities:[44]

$$\begin{aligned} P(\text{V.3sg}|\text{SUBJ}) &= 0.651 \\ P(\text{V.3sg}|\text{OBJ}) &= 0.606 \\ P(\text{V.3pl}|\text{OBJ}) &= 0.152 \end{aligned}$$

[43] If we did not assume independence of the verb form from the lexical content of the initial noun phrase, the effect would most likely be to increase of the conditional probability of *unterstützt* relative to *unterstützen*, because the former is one of presumably a rather narrow range of semantically plausible verbs given *System* as the grammatical subject, whereas *System* as grammatical object is semantically compatible with a wide range of transitive verbs.

[44] Note that nearly all the verbs in the corpus are third-person.

The crucial comparison is between the second and third lines: even when the initial NP is an object, the next word is far more often a singular finite verb than a plural finite verb.[45] With these probabilities we can now estimate the expectations for singular and plural finite verbs:

$$\begin{aligned} P(\text{V.3sg}|\text{WS}) &= 0.45 \times 0.651 + 0.55 \times 0.606 \\ &= 0.626 \\ P(\text{V.3pl}|\text{WS}) &= 0.55 \times 0.152 \\ &= 0.0836 \end{aligned}$$

The resulting probability ratio, $7.5 : 1$ in favor of singular finite verbs, outweighs the $2.3 : 1$ ratio we estimated for the probability of the verb form given the number-marked part of speech. Therefore, the surprisal at *unterstützt* is less than the surprisal at *unterstützen*, which is consistent with empirical reading-time results—even if the relevant initial NPs are in fact more likely to be objects than subjects.

[45] Neither set of conditional probabilities sums to 1 because the next word following the sentence-initial NP may not yet be the finite verb.