

Memory and surprisal in human sentence comprehension

Roger Levy

31 October 2012

INTRODUCTION

Humboldt famously described language as a system of rules which “makes infinite use of finite means” (Humboldt, 1836; Chomsky, 1965) and this is doubly true in the study of language comprehension. On the one hand, the comprehender’s *knowledge* of language must be finitely characterized: the brain itself as a computational device is finite, as is the comprehender’s experience of her native language. Hence understanding is an act of generalization: the comprehender must apply the knowledge gleaned from her lifetime of previous linguistic experience to analyze a new sentence and infer what the speaker is likely to have intended to mean. This need for analysis gives rise to the second sense in which Humboldt’s aphorism is true: to understand a sentence in real time the comprehender must deploy her limited cognitive resources to analyze input that is potentially unbounded in its complexity. Nowhere are these truths more evident than in the determination of sentence structure during language comprehension, as in (1) below. Before you go on reading, take as much time as you need to fully understand this sentence, and while you are doing so, reflect upon what you find difficult about it. You may even want to write down your reflections on its difficulty so that you can remind yourself of them once you reach the end of the chapter.

- (1) Because the girl that the teacher of the class admired didn’t call her mother was concerned.

I am confident that you have never encountered this sentence before, but you probably understood it fully with some effort. In order to understand it, you had to correctly construct all the structural relationships it contains: that *girl* is both the subject of *didn’t call* and the object of *admired*, that the subject of *admired* is *teacher*, that there is a clause boundary between *call* and *her mother* and thus that *her mother* is the subject of *was* but not the object of *admired*, and so forth. You have probably encountered few if any sentences with the precise array of structural relationships seen here, but the individual elements are familiar; your ability to understand the sentence at all rests on your ability to put these elements together in novel configurations in real time, despite the occasional difficulty involved.

As this sentence illustrates, although we are generally successful (perhaps remarkably so) in our ability to achieve linguistic understanding in real time, hallmarks of our limited experience and cognitive resources do exist and can be measured: misunderstanding does

occur, and even among sentences that are successfully understood, difficulty is differential and localized. That is, not all sentences are equally easy to understand, nor are all parts of a given sentence equally easy to get through. There are two places where you probably found sentence (1) especially difficult: around the word *admired*, where you probably felt uncomfortable with having to keep track of the relationships among the preceding elements *girl*, *teacher*, and *class*; and at the phrase *was concerned*, where you probably were taken aback at encountering the main verb of the sentence without prior warning. By the time we reach the end of this chapter, you will have learned about leading theories of real-time sentence comprehension that account both for your ability to understand sentence (1) and for these sources of difficulty that you may have experienced in doing so.

This chapter thus presents a broad outline of two approaches to understanding these cognitive underpinnings of real-time language comprehension. Each approach is rooted in a deep intuition regarding the nature of limitations in the cognitive resources deployed during sentence understanding. One focuses on *memory*—the use of cognitive resources for storage and retrieval of the representational units used in analysis of linguistic input. The other focuses on *expectations*—the pre-emptive allocation of cognitive resources to various alternatives in the face of uncertainty. In each case, the hypothesis is that the resources in question are sufficiently scarce so as to form a *bottleneck* in real-time comprehension: overtaxing these resources, either by overloading memory or by presenting the processor with a sufficiently unexpected event, can create measurable disruption in real-time comprehension. Each approach has a rich history in the literature, has been formalized mathematically, and enjoys considerable empirical support. Yet there are cases where the two come into conflict, and their proper resolution remains to be fully understood. In this chapter I begin with memory-based approaches, continue with expectation-based approaches, and then turn to cases where the two come into conflict.

MEMORY LIMITATIONS, LOCALITY, AND INTERFERENCE

The traditional picture of memory limitation in sentence comprehension

The notion of limited memory as a bottleneck on language comprehension dates back to the earliest days of modern psycholinguistics. In the late 1950s, Chomsky introduced the *competence/performance* distinction of knowledge versus patterns of usage of a language, and with it introduced phrase-structure grammars as a formal means of characterizing key aspects of a native speaker's syntactic competence (Chomsky, 1956, 1957). It was immediately recognized, however, that a wide range of sentences generated by linguistically plausible competence grammars could not actually be understood by native speakers, raising questions regarding the relationship between the competence grammars posited by the new generative linguistics and actual linguistic performance. One answer to this problem put forward by George Miller and others (Miller and Chomsky, 1963; Miller and Isard, 1963; Marks and Miller, 1964) was the strong position that competence grammars were essentially faithful characterizations of a speaker's psychological knowledge, but that *performance constraints* interfered with the effective deployment of this competence for some types of sentences. One

such type was the *multiply center-embedded* sentence, such as (2) below (Yngve, 1960; Miller and Chomsky, 1963):

- (2) This is the malt that the rat that the cat that the dog worried killed ate.

This sentence is simply generated by the repeated application of grammatical rules for forming object-extracted relative clauses, as in (3) below:

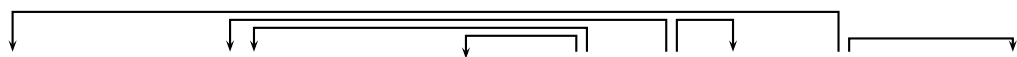
- (3) *the dog* worried the cat \Rightarrow the cat that *the dog* worried
the cat that the dog worried killed the rat \Rightarrow the rat that *the cat that the dog worried*
 killed

Despite the straightforwardness of its derivation, however, (2) is extremely difficult to comprehend. That this difficulty cannot be ascribed purely to the complexity of the meaning of the sentence can be seen by comparison with (4) below, which is essentially synonymous but much easier to understand.

- (4) This is the malt that was eaten by the rat that was killed by the cat that was worried by the dog.

The earliest work on this problem (Yngve, 1960; Miller and Chomsky, 1963; Chomsky and Miller, 1963) attributed the difficulty of (2) to the large number of incomplete and nested syntactic relationships that must be maintained partway through the string. Figure 1 illustrates the situation for (2), assuming that the comprehender’s incremental representation of sentence structure is captured by a left-to-right incrementally expanded context-free tree. After the final instance of *the*, the sentence has reached a fourth level of center-embedding, and a stack of four categories must be kept in memory for faithful completion of the tree when further input is encountered. Yngve (1960) proposed a model in which human incremental language comprehension assigns such incremental structural representations but has severely limited (3 or less) stack depth, making complex center-embedded sentences incomprehensible.

However, it soon became clear that such a straightforward characterization of memory limitation was unworkable (Miller and Chomsky, 1963; Gibson, 1991). For one thing, it is empirically the case that processing breakdown seems to happen when the comprehender *emerges* from the deeply center-embedded structure, around the word *killed*, rather than when the comprehender *arrives* at the deepest embedding level (Gibson and Thomas, 1999; Christiansen and MacDonald, 2009; Vasishth et al., 2010). Second, Yngve’s strictly top-down model had no good mechanism for explaining how recursively left-branching structures, prevalent in head-final languages such Japanese, are parsed. Third, the *type* and *arrangement* of embedding structures turns out to matter as much as the sheer depth of embedding. A particularly clear example of this latter point can be seen in the contrast of (5) below (the verb-argument dependency arrows drawn in the example will be used later on to describe memory-based analyses of processing difficulty for these examples):

- (5) a. 
 The fact that the bike messenger who the car just missed blocked traffic angered the motorcyclist.

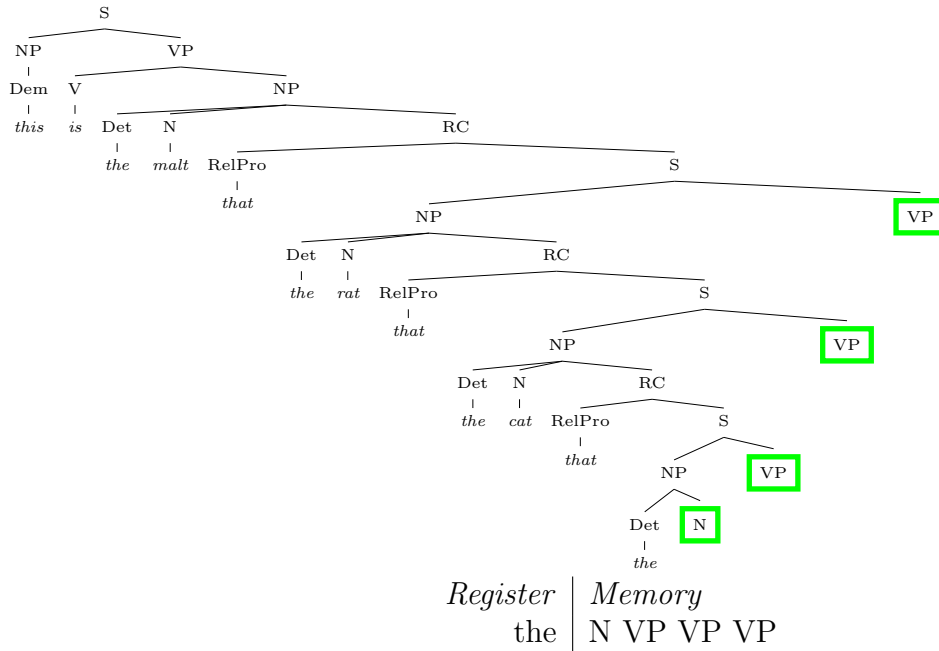
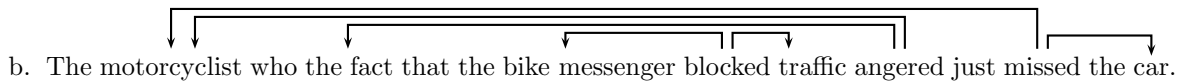


Figure 1: Deep inside a multiply center-embedded sentence in the stack-depth model of Yngve (1960)



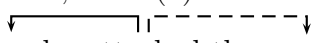
In (5a), an object-extracted relative clause (*who the car just missed*; ORC) is embedded inside a complement clause (*that the bike messenger who the car just missed blocked traffic*; CC). We have the reverse situation in (5b), where a CC is embedded inside an ORC. Gibson and Thomas (1997; see also Cowper, 1976; Gibson, 1991) demonstrated that the CC-inside-ORC variant (5b) is considerably harder to understand than the ORC-inside-CC variant (5a), despite the fact that the depth of phrase-structural center-embedding—two levels—is identical in these cases.

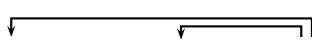
Observations such as these have drawn particular attention to the moments during syntactic comprehension when links between words can be constructed establishing particular aspects of sentence meaning. The key difference between noun-modifying relative and complement clauses is that whereas the former simply involve modification of a noun by a clause that could itself stand alone as an independent sentence, the latter involve EXTRACTION of an NP, such that proper interpretation requires reconstruction of the relationship between the head noun and the element governing the extraction site (Chomsky, 1981; Pollard and Sag, 1994). Intuitively, an underlying verb-object DEPENDENCY relation needs to be established between *missed* and *bike messenger* in (5a) and between *angered* and *motorcyclist* in (5b), but not between *blocked* and *fact* in either example. Notably, the linear distance between the verb and the object resulting from the extraction is considerably greater for the CC-inside-RC example (*angered*→*motorcyclist* in (5b)) than for the RC-inside-CC example (*missed*→*bike messenger* in (5a)). Furthermore, more noun phrases intervene between the

RC verb and its object in (5b) than in (5a).

Two prominent theories of memory in online syntactic comprehension have arisen from data of the type described above. One theory is the DEPENDENCY LOCALITY THEORY (DLT, also called Syntactic Prediction Locality Theory; Gibson, 1998, 2000; Grodner and Gibson, 2005), which deals with the integrations between elements in a sentence (in practice, usually word-word dependencies), as well as with expectations for those integrations. In DLT, there are two types of costs that can be incurred during processing of part of a sentence: the INTEGRATION cost incurred during the establishment of the dependencies between the currently-processed word(s) and earlier parts of the sentence; and the STORAGE cost incurred for maintaining representations of incomplete dependencies that will be completed later in the sentence. In the theory, integration costs are larger (i) the more new dependencies are constructed at once, and (ii) the more material intervening between the governor and governed elements of each dependency. Hence DLT is able to capture the pattern seen in (5). At the word *angered* in (5b), two dependencies must be constructed simultaneously: one between *angered* and its object, *motorcyclist* (with three intervening nouns—*fact*, *bike messenger*, and *traffic*), and another between *angered* and its subject, *fact* (with two intervening nouns—*bike messenger* and *traffic*). In (5a), there is no corresponding word requiring such a complex set of simultaneous integrations: at each verb where two integrations are required, at least one of the governed NPs is linearly adjacent.

A range of other evidence has also been adduced in support of the integration component of DLT, particularly from the comprehension of different types of relative clauses (e.g., Warren and Gibson, 2002; Hsiao and Gibson, 2003; Gordon et al., 2004, and see below). One of the best-known examples is the asymmetry in comprehension difficulty of subject-extracted versus object-extracted RCs in English when both the head noun and the RC-internal NP are animate, as in (6) below:

- (6) a. The reporter who attacked the senator left the room.


 b. The reporter who the senator attacked left the room.


The integration cost at the RC verb *attacked* is greater in the ORC (6b) than in the SRC (6a), since in the ORC there are two preceding dependents that must simultaneously be integrated, one of which is not adjacent to the verb; whereas in the SRC the dependents are integrated one by one—in (6a), the dotted line reflects that when *attacked* is processed *the senator* has not yet been seen and thus is not yet integrated—and each is adjacent to the verb. The greater processing difficulty of such ORCs has been demonstrated empirically in many studies (Wanner and Maratsos, 1978; Ford, 1983; King and Just, 1991; Gordon et al., 2001; Grodner and Gibson, 2005, inter alia). Evidence for the storage component is scarcer, though see Chen et al. (2005); Gibson et al. (2005); Grodner et al. (2002); Nakatani and Gibson (2010) for published studies. We will return to the study of Grodner et al. (2002) shortly.

The second theory has many qualitative similarities to DLT, but differs in focusing more exclusively on the integration component of comprehension, and in placing greater emphasis

on grounding the content of integration operations in influential theories of working memory within cognitive psychology, specifically theories of CONTENT-ADDRESSABLE memory with CUE-BASED RECALL. In the most explicit instantiations of such a theory (Lewis and Vasishth, 2005; Lewis et al., 2006), the parser has no explicit representation of linear order of preceding input, including the relative priority of preceding sites of potential syntactic attachment (contrasting with theories such as Yngve’s; Figure 1). Rather, an incremental, tree-structured representation of the sentence thus far is stored in content-addressable memory, and in order to construct new dependencies between the currently processed word and preceding content, the current word serves as a cue for recall of the appropriate integration site(s). In the SRC of (6a) above, for example, upon reading the word *attacked* the parser must retrieve the representation of *reporter* from content-addressable memory in order to link it as the argument of the current word.

One crucial component of the cue-based recall theory is that representations of all preceding syntactic items stored in working memory compete with one another during retrieval. Two factors affect the ease of retrieval of the correct (TARGET) unit. First, retrieval is easier the better the match between the features of the cue and target, relative to the degree of match between the features of the cue and other, non-target, items in working memory. Second, in some variants (Lewis and Vasishth, 2005; Lewis et al., 2006), retrieval is easier the greater the ACTIVATION LEVEL of the target item relative to the activation level of other, non-target items plays a role: items have a high activation level when first encoded in memory, that activation decays over time, but every retrieval boosts the item’s activation. The theory thus predicts the same differential difficulty effect observed in the English SRC/ORC contrast of (6) and predicted by the DLT. In (6b), as in (6a), one of the operations that has to take place at the RC verb *attacked* is retrieval of the representation of the head noun *reporter* and construction of a dependency of the appropriate type between it and the verb. In (6b), however, unlike in (6a), another noun—*senator*—is already encoded in memory, and successful retrieval requires correct association of these nouns with the subject and object roles for *attacked* respectively. The semantic similarity of the two nouns leads to high retrieval interference, slowing processing and even creating the possibility of misretrieval—that is, interpreting the RC as semantically reversed (as *the reporter attacked the senator*). The predictions of the cue-based recall theory in terms of online processing effects thus match those of the DLT, with greater reading times predicted at *attacked* in (6b) than in (6a).

Predictions of the cue-based theory and the DLT diverge, however, for cases such as (7) below:

(7) The movie that the director watched received a prize.

For the DLT, the RC verbs in (7) and (6b) incur exactly the same integration costs, since they both involve integrating two dependents with the same distances. In the cue-based recall theory, however, two factors make argument retrieval at the RC verb easier in (7) than in (6b): first, *movie* and *director* are less semantically similar than *reporter* and *senator*, making their memory representations more distinct; second, the properties of *movie* do not match the retrieval cues for the subject position of *watched*, since only animate entities can

perform watching. This difference in predicted processing difficulty was confirmed by Traxler et al. (2002).

Applications beyond center-embedding difficulty

Although the study of memory limitation in online sentence comprehension has its roots in processing difficulty and breakdown effects associated with unambiguous center-embedding and retrieval difficulty, the resulting theories have been applied to a considerably wider variety of phenomena. Gibson (1991), for example, introduced the idea that syntactic ambiguity resolution in online comprehension might attempt to minimize memory storage costs due to unfulfilled syntactic expectations of the sort encoded in the DLT. One study exploring such an idea is Grodner et al. (2002), who examined sentences like (8) below.

- (8) a. The river which the valley (that was) captured by the enemy contains has its source at a glacier. [RC]
b. The commander knows that the valley (that was) captured by the enemy contains a river. [SC]

According to DLT in both cases, when *the valley* is processed there is an expectation generated for an upcoming verb for which it is the subject. Further downstream, when the words *that was* are absent (the AMBIGUOUS variants) there is a temporary syntactic ambiguity at *captured* between a finite-verb interpretation (e.g., *The river which the valley captured the sunlight reflecting off of was flowing quickly*) and a reduced-relative interpretation (as in (8a)). On the finite-verb interpretation, *captured* completes the upcoming-verb expectation generated earlier, but on the reduced-relative interpretation this expectation remains unmet. Hence the reduced-relative interpretation imposes a higher memory-storage cost than the finite-verb interpretation. But as we already saw in analysis of (5), the RC context of (8a) itself is more memory-intensive than the SC context of (8b). Thus, if comprehenders avoid especially memory-intensive interpretations—as the RC context combined with the reduced-relative interpretation would lead to—the finite-verb interpretation should be more strongly preferred in the ambiguous variant of (8a) than in the ambiguous variant of (8b). Indeed, Grodner et al. found an interaction between ambiguity and embedded-clause type at the critical region *by the enemy*, which disambiguates the structure toward a reduced relative interpretation; reading times were superadditively greatest in the ambiguous RC condition, suggesting that the RC context may indeed induce comprehenders to entertain a finite-verb analysis of *captured* (though see further discussion under *Conflicting predictions between expectations and memory*).

Let us turn now to interference-based theories and a distinctive type of prediction they make, involving cases where preceding context may make retrieval of preceding dependents at a verb not only *difficult* but even *inaccurate*. Here I briefly outline two examples. Wagers et al. (2009) have applied interference-based theory in the study of AGREEMENT ATTRACTION in online comprehension. To explain the phenomenon of agreement attraction, which has been studied primarily in the sentence production literature, consider example (9) below.

- (9) a. The key to the cabinets were rusty from many years of disuse. [UNGRAMMATICAL, +ATTRACTOR]
 b. The key to the cabinet were rusty from many years of disuse. [UNGRAMMATICAL, -ATTRACTOR]

Each of these sentences contains an agreement “error” in which the number marking on the finite verb *were* fails to match the number of the subject NP, which is determined by the head noun of the subject, in this case *key*. Agreement attraction is the phenomenon of errors of the type in sentence (9a), where the verb’s number marking matches the number of some noun (the ATTRACTOR, here *cabinets*) other than the true head of the subject, being more common than errors of the type in sentence (9b) (Bock and Miller, 1991; Eberhard, 1999; Eberhard et al., 2005; Franck et al., 2002; Vigliocco and Nicol, 1998, inter alia). One of the leading theories from the field of language production is that attraction effects arise from PERCOLATION of agreement features from within a complex NP (here, plural number from *cabinets*) up to the NP head, leading to incorrect representation of the subject NP’s number.

In the comprehension literature, Pearlmutter et al. (1999) had previously found that plural attractors effectively weakened the precision of comprehenders’ online assessment of subject-verb agreement. Reading times immediately after the verb in both (9a) and (9b) are inflated compared with singular-verb variants (*was* instead of *were*), but plural attractors reduce the reading-time penalty. The percolation theory of the sentence-production literature can explain the results of Pearlmutter et al. in comprehension: if the comprehender misrepresents the number of the complex NP *the key to the cabinets*, then this might lead to failure to identify the agreement anomaly when the correct syntactic relationship between *were* and the preceding subject NP is constructed. Wagers et al., however, showed that attraction can equally affect comprehension of verbs inside relative clauses, using sentences such as (10) below:

- (10) a. The musician who the reviewer praise so highly will probably win a Grammy.
 b. The musicians who the reviewer praise so highly will probably win a Grammy.

Reading times immediately after the RC verb *praise* in (10b) are deflated relative to those in (10a), suggesting that plural marking on the RC head noun can reduce the strength of the anomaly experienced at a singular verb whose subject is an RC-internal singular NP (a similar acceptability pattern was first reported by Kimball and Aissen, 1971). This result is not explained by the percolation theory, because the plural noun (*musicians*) is not inside the RC subject (*the reviewer*) and thus upward percolation of the plural feature would not lead to an incorrect representation of the RC subject’s number. As Wagers et al. describe, in an interference-based framework this pattern could arise from the possibility of incorrect retrieval of the RC head for the subject slot of the RC verb, which would result in failure to detect the subject-verb number mismatch. That is, under the interpretation of Wagers et al. the number mismatch may lead to an *incorrect syntactic relationship* being entertained or even established between *praise* and *reviewers*. Although these results cannot themselves adjudicate completely between theories in which the true subject’s number features are incor-

rectly represented and theories in which the wrong NP is retrieved for the verb's subject slot, they speak to the ability of interference-based theories to make testable predictions regarding online comprehension difficulty ranging over a wide variety of syntactic configurations.¹

A second example relates to the understanding of LOCAL COHERENCE EFFECTS, where a grammatical analysis that would be available for a substring of a sentence only when that substring is taken in isolation seems to compete with the global grammatical analysis:

- (12) a. The coach smiled at the player *tossed* the frisbee.
b. The coach smiled at the player *thrown* the frisbee.

In (12a), *the player tossed* would be analyzable in isolation as the onset of an independent clause with *the player* as the subject and *tossed* as the main verb, but that analysis is inconsistent with the grammatical context set up by the rest of the sentence. This is not an issue in (12b) as the *thrown* does not have the part-of-speech ambiguity that allows it to serve as a finite verb. Tabor et al. (2004) showed that reading times were greater starting at this critical verb in sentences like (12a) as compared with (12b), and argued for a model in which bottom-up and top-down syntactic analysis took place simultaneously and could thus come into conflict (see also Bicknell and Levy (2009) for a Bayesian variant of such a model).²

Van Dyke (2007), however, points out that an interference-based model such as that of Lewis and Vasishth (2005) can accommodate local-coherence effects, since *the player* might sometimes be incorrectly picked out by the subject-retrieval cues of *tossed*. Van Dyke goes on to show examples where a match between a verb's retrieval cues and an NP that is not immediately adjacent can induce similar processing difficulty, as in (13) (see also Van Dyke and Lewis, 2003 for related studies):

- (13) The worker was surprised that the resident who said that *the warehouse/neighbor* was dangerous was complaining about the investigation.

Here, differential processing difficulty begins at the region *was complaining*, with greater difficulty when the sentence contains *the neighbor* than when it contains *the warehouse*. This result suggests that when the preceding NP matches the semantic requirements made by the main-clause verb on its subject, it is sometimes entertained as the subject of the main-

¹As Roger van Gompel points out, interference-based theories make an incorrect prediction for the pattern of verbal agreement processing for *grammatical* sentences as in (11) below:

- (11) a. The key to the cabinet was rusty from many years of disuse. [GRAMMATICAL, -ATTRACTOR]
b. The key to the cabinets was rusty from many years of disuse. [GRAMMATICAL, +ATTRACTOR]

Interference-based theories predict that the verb should be more difficult when both nouns inside the subject NP are singular, as in (11a), since both nouns match the verb's target cue, whereas in (11b) only the true subject matches this cue. But no trace of this pattern was found by either Pearlmutter et al. (1999) or Wagers et al. (2009).

²Although the local-coherence effect seems difficult to accommodate in a model where syntactic analysis is entirely top-down, Levy (2008b) presents such a model in which the effect arises from uncertainty about the representation of preceding context; Levy et al. (2009) confirm predictions made by this model.

clause verb even though neither the global nor the local syntactic context would license such an analysis.

EXPECTATION-BASED COMPREHENSION AND SURPRISAL

Equally fundamental as the intuition that memory limitations affect online sentence comprehension is the intuition that a language user's context-derived EXPECTATIONS regarding how a sentence may continue can dramatically affect how language comprehension unfolds in real time. Among the best-known early demonstrations of this phenomenon are the SHADOWING studies pioneered by Marslen-Wilson (1975), who demonstrated that listeners continuously repeating back speech they hear with lags as short as a quarter-second are biased to correct disrupted words; for example, when shadowing *He's departing the day after tomorrane* the listener might correct the final word to *tomorrow*, but only when the corrected form of the word was syntactically and semantically consistent with context. This result indicated the extreme rapidity with which comprehenders use context to constrain the interpretation of new linguistic input—in this case, recognition of a word's identity.

Since then, the known empirical scope of this biasing effect of context-derived expectation has expanded in two key respects. First, it is now known that correct expectations increase the *rate* at which novel input is processed. Ehrlich and Rayner (1981) demonstrated that words which are strongly predicted by their preceding context, as measured by the “fill in the blank” Cloze completion method (Taylor, 1953), are read more quickly than unpredictable words. Hence, of the two contexts

- (14) a. The boat passed easily under the ___
b. Rita slowly walked down the shaky ___

the word strongly predicted for context (14a) is read more quickly in that context than in (14b).³ Analogous signatures of correctly matched expectations can also be found in EEG responses during online sentence comprehension (Kutas and Hillyard, 1980, 1984; see also Van Berkum et al., 2005; DeLong et al., 2005, and Wicha et al., 2004 for more recent evidence). Second, it is now known that incremental discrimination among alternative analyses of structurally ambiguous input is exquisitely sensitive to (both linguistic and non-linguistic context). To take a well-known example, as the sentence onset *Put the apple on the towel. . .* is uttered, the listener's interpretation of *on the towel* (is it describing which apple to move, or where to put the apple?) is strongly influenced by how many apples are present in a visible physical array (Tanenhaus et al., 1995; see also Altmann and Kamide, 1999; Trueswell et al., 1994; MacDonald, 1993, among others; see also Spivey, McRae, and Anderson, this volume). In the 1990s, two classes of computational models were proposed which had a lot to say about how diverse information sources influenced ambiguity resolution: CONSTRAINT-BASED models inspired by neural networks (Spivey and Tanenhaus, 1998; McRae et al., 1998; Tabor and Tanenhaus, 1999; see also McRae and Matsuki, this volume) and PROBABILISTIC

³You have probably already guessed: the vast majority of native English speakers fill in the blank with the word *bridge* (Arcuri et al., 2001).

GRAMMAR-BASED DISAMBIGUATION models (Jurafsky, 1996; Narayanan and Jurafsky, 1998, 2002; Crocker and Brants, 2000). Because these models covered only resolution of ambiguity in the grammatical analysis of input that had already been seen, however, they had little to say about expectation-derived processing speedups in examples like (14) above, or about the rich set of syntactic-complexity effects found in what are for the most part structurally unambiguous situations (see section on *Memory limitations, locality, and interference*).

Surprisal

In 2001, however, Hale, drawing inspiration from Attneave (1959), proposed a quantification of the cognitive effort required to process a word in a sentence—the SURPRISAL of the word in the context it appears—which has raised prospects for a unified treatment of structural ambiguity resolution and prediction-derived processing benefits. Surprisal (sometimes called “Shannon information content” in the information theory literature) is defined simply as the log of the inverse of the probability of an event; in the case of a word w_i in a sentence following words w_1, \dots, w_{i-1} and in extra-sentential context C , the surprisal is thus simply

$$\log \frac{1}{P(w_i|w_1, \dots, w_{i-1}, C)} \quad (1)$$

Hale focused on the framing of incremental sentence comprehension as the step-by-step disconfirmation of possible phrase-structural analyses for the sentence, leading to an interpretation of the cognitive load imposed by a word as “the combined difficulty of disconfirming all disconfirmable structures at a given word”. On that view, surprisal emerges as a natural metric of word-by-word cognitive load on the assumption that more probable structures are more work to disconfirm (see section on *Theoretical Justifications for Surprisal* for greater discussion of this assumption).

Surprisal and garden-path disambiguation

Hale (2001) and Levy (2008a) cover a range of psycholinguistic phenomena which can be successfully analyzed within the surprisal framework ranging from classic instances of garden-path disambiguation (*the horse raced past the barn fell*; Bever, 1970) to processing benefits when ambiguity is left unresolved (Traxler et al., 1998; van Gompel et al., 2001, 2005) to syntactic-expectation-based facilitation in unambiguous contexts (see section on *Constrained syntactic contexts*). To give the reader a more concrete picture of how surprisal can simultaneously account for both empirically observed syntactic-processing effects which involve disambiguation and effects which do not, I provide here a novel and fairly explicit illustration of how probabilistic grammatical analysis can be combined with surprisal to derive predictions for a well-studied construction which turns out to exhibit both types of effects. Example (15) from Staub (2007) below serves as a starting point:

(15) When the dog scratched the vet and his new assistant removed the muzzle. ⁴

Garden-path disambiguation is an important feature of this sentence: the phrase *the vet and his new assistant* creates a temporary structural ambiguity: this phrase could be the object NP of the subordinate-clause verb *scratched*; or it could be the subject NP of the main clause, in which case *scratched* would have no overt object (Frazier and Rayner, 1982; Clifton, 1993; Ferreira and Henderson, 1990; Mitchell, 1987; Adams et al., 1998; van Gompel and Pickering, 2001; Pickering and Traxler, 1998; Sturt et al., 1999; Staub, 2007). Intuitively, the preferred initial interpretation is as the object of *scratched*, which is globally incorrect; the strongest disambiguatory evidence comes at the main-clause verb *removed*. The measurable real-time correlate of this disambiguation effect was first demonstrated by Frazier and Rayner (1982), who showed that the amount of time that the eyes linger upon the disambiguating material is elevated in cases like (15) when compared with cases such as (16) below, in which the presence of either an overt NP object of *scratched* (*its owner* in (16a)) or a comma marking the end of the subordinate clause (16b) facilitates the initial interpretation of the following NP as the main-clause subject:

- (16) a. When the dog scratched its owner the vet and his new assistant removed the muzzle.
 b. When the dog scratched, the vet and his new assistant removed the muzzle.

How can the relative difficulty of (15) be captured in a framework of sentence comprehension as probabilistic grammar-based inference? Table 1 illustrates a small PROBABILISTIC CONTEXT-FREE GRAMMAR (PCFG; Booth, 1969; Manning and Schütze, 1999) which includes the grammatical rules necessary to cover both the garden-path and globally correct interpretations of examples (15) and (16). Intuitively, a PCFG both states what grammatical structures are *possible* (determined by the set of rules in the grammar) and distinguishes the relative likelihood of different possible grammatical structures (with more likely grammatical structures given higher probability values).⁵ The syntactic categories and style of phrase structure rule used here are chosen to roughly conform to those used in the Penn Treebank (Marcus et al., 1994), the most widely used syntactically annotated corpus in computational linguistics. The probabilities in this grammar are chosen by hand for expository purposes, but they reflect two important facts about the distributions of the relevant constructions in naturalistic English text: first, verb phrases can be either transitive or intransitive (reflected in the presence of both $VP \rightarrow V NP$ and $VP \rightarrow V$ rules); second, most but not all sentence-initial subordinate clauses are delimited on the right by a comma (reflected in the

⁴To keep the grammar used for exposition small, I have substituted *removed* for the phrase *took off* actually used by Staub, to avoid verb-particle constructions. None of the analyses are qualitatively affected by this change.

⁵More technically, a PCFG is a collection of context-free grammatical rules of the form $X \rightarrow \alpha$, where X is a single non-terminal symbol (syntactic category) and α is a sequence of symbols (syntactic categories and/or words), each of which has a probability. The probabilities are constrained such that for every non-terminal symbol X in the grammar, the probabilities of all rules with X on the left-hand side sum to 1: $\sum_{\alpha} P(X \rightarrow \alpha) = 1$. The product of a tree T is the product of the probability of each rule used in the derivation of T (if a rule is used more than once, its probability is multiplied in each time it's used), and the product of a string $w_{1\dots n}$ is the sum of the products of all trees whose yield (the leaves of the tree, read from right to left) equals $w_{1\dots n}$. The interested reader is encouraged to consult Jurafsky and Martin (2008) or Manning and Schütze (1999) for more details.

high probability of the SBAR → COMPL S COMMA rule).

To understand how probabilistic syntactic knowledge and incremental comprehension interact to yield garden-path disambiguation effects within the surprisal framework, let us consider the probability distribution over incremental parses of the pre-disambiguation sentence prefixes shown in (17) below:

- (17) a. When the dog scratched the vet and his new assistant...
b. When the dog scratched *its owner* the vet and his new assistant...

In (17a), the PCFG of Table 1 makes two incremental parses available, shown in Figure 2. These trees are “incremental” in the sense that, aside from nodes strictly dominating input that have already been seen, only nodes that can be inferred with probability 1 are indicated. To explain how inference about a sentence’s syntactic structure arises from the application of probabilistic grammars, we introduce a bit of notation: let $w_{1\dots i}$ denote the sentence thus far (up to the current word w_i) and let the variable T denote some incremental syntactic parse that is logically possible given the grammar and the sentence thus far. The probability of each incremental parse T can be computed using Bayes’ Rule, a basic theorem of probability theory, according to which we can say that

$$P(T|w_{1\dots i}) = \frac{P(T)}{P(w_{1\dots i})} \quad (2)$$

That is, the probability of a particular parse T given the string observed thus far is equal to the probability assigned to the parse by the grammar, divided by the total probability of the partial sentence seen thus far.⁶ Thus, only parses consistent with the grammar are permitted, and among them, those with higher probability given by the grammar are preferred over those with lower probability. When these computations are applied to the examples in Figure 2, we find that the tree in which *the vet and his new assistant* is interpreted as the direct object of *scratched* has probability 0.826, and the tree in which it is interpreted as the main-clause subject has probability 0.174.⁷ In (17b), by contrast, only the main-clause subject interpretation is available (with probability 1) for *the vet and his new assistant*, due to the

⁶To be more technically precise, Bayes Rule tells us that

$$P(T|w_{1\dots i}) = \frac{P(w_{1\dots i}|T)P(T)}{P(w_{1\dots i})}$$

but by definition $P(w_{1\dots i}|T)$ is 1 when the yield of T is $w_{1\dots i}$ and 0 otherwise, so if we limit ourselves to considering parses consistent with the input sentence we can just drop the first term in the numerator, giving us (2).

⁷This difference in conditional probability between the two analyses arises from three differences between the two incremental trees: namely, in the main-clause analysis:

1. the subordinate-clause VP rewrites to V rather than to V NP;
2. there is a commitment that only one subordinate-clause SBAR node initiates the sentence;
3. there is a commitment that there is no comma between the subordinate clause and the main clause.

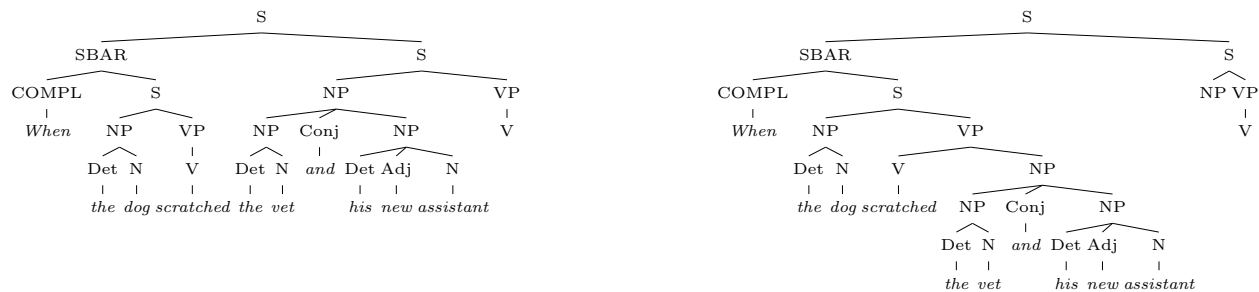


Figure 2: The two incremental analyses for sentence (15) pre-disambiguation.

presence of *its owner* as direct object of *scratched*. With respect to the probabilities assigned to incremental interpretations of a sentence, surprisal theory is thus quite similar to pruning and attention-shift theories of garden-path disambiguation Jurafsky (1996), Narayanan and Jurafsky (1998, 2002), and Crocker and Brants (2000).

Within the simplest version of surprisal theory, however, the garden-path disambiguation effect itself arises not from complete loss of the correct analysis but from the comprehender’s need to hedge her predictive bets regarding how the sentence may continue. Let us ask: given that after processing the word *assistant* the two structures of Figure 2 are maintained with the probabilities just stated, how likely is it that the next word of the sentence is the word *removed*? According to the laws of probability theory, both structures contribute to predicting how likely *removed* is to be the next word in the sentence, but the more likely structure plays a larger role in determining the strength of the prediction. Under the main-clause subject analysis, the conditional probability of *removed* being the next word is 0.2;⁸ under the direct-object analysis, the conditional probability is 0 since a verb cannot appear until after a main-clause subject has been encountered. The surprisal of *removed* is simply the weighted average of these two probabilities:⁹

$$P(w_{11} = \textit{removed} | w_{1...10}) = 0.2 \times 0.174 + 0 \times 0.826 = 0.0348$$

so the surprisal is $\log_2 0.0348 = 4.85$ bits. The corresponding surprisal for (17b), in which the incremental syntactic analysis was unambiguously main-clause subject, is $\log_2 0.2 = 2.32$ bits (Table 2). Hence surprisal theory correctly predicts the difference in processing difficulty due to this case of garden-pathing.

It is worth noting in this example grammar that no distinction is made between transitive and intransitive verbs. However, Mitchell (1987) and van Gompel and Pickering (2001) (see

⁸This conditional probability reflects (i) that the subject must not continue with another NP conjunct; and (ii) that the main-clause verb must turn out to be *removed*.

⁹In probability theory, the determination of this weighted average is called MARGINALIZATION; in its general form for this example we would write that:

$$P(w_{11} = \textit{removed} | w_{1...10}) = \sum_T P(w_{11} = \textit{removed} | w_{1...10}, T) P(T | w_{1...10})$$

and we would say that the probability of the upcoming word is computed “marginalizing over the possible structural analyses of the sentence thus far”.

	Rule	Prob.		Rule	Prob.		Rule	Prob.
S	→ SBAR S	0.3	Conj	→ and	1	Adj	→ new	1
S	→ NP VP	0.7	Det	→ the	0.8	VP	→ V NP	0.5
SBAR	→ COMPL S	0.3	Det	→ its	0.1	VP	→ V	0.5
SBAR	→ COMPL S COMMA	0.7	Det	→ his	0.1	V	→ scratched	0.25
COMPL	→ When	1	N	→ dog	0.2	V	→ removed	0.25
NP	→ Det N	0.6	N	→ vet	0.2	V	→ arrived	0.5
NP	→ Det Adj N	0.2	N	→ assistant	0.2	COMMA	→ ,	1
NP	→ NP Conj NP	0.2	N	→ muzzle	0.2			
			N	→ owner	0.2			

Table 1: A small PCFG for the sentences in section on *surprisal and garden-path disambiguation*

also Staub, 2007) provided relevant evidence evidence by comparing reading of sentences like (15) with sentences like (18) with intransitive subordinate-clause verbs.

(18) When the dog arrived the vet and his new assistant removed the muzzle.

These studies in fact revealed *two* interesting effects. First, early reading times at the main-clause verb (*removed* in this case) were elevated for transitive as compared with intransitive subordinate-clause verb sentences. This is precisely the effect predicted by incremental disambiguation models in which fine-grained information sources are used rapidly: before encountering the main-clause verb, the comprehender is already much more committed to a main-clause analysis when the subordinate-clause verb is intransitive. Equally interesting, however, early reading times at the onset of the potentially ambiguous NP (*the vet* in this case) were *lower* for transitive as compared with intransitive subordinate-clause verb sentences. This effect is *not* obviously predicted by all incremental disambiguation models using fine-grained information sources; the constraint-based model of Spivey and Tanenhaus (1998) and McRae et al. (1998), for example, which predict processing slow-downs when a structural ambiguity is encountered and relative preferences for the alternative interpretations need to be determined, might well predict *greater* difficulty at the ambiguous-NP onset in the transitive case, since there is a true structural ambiguity only when the preceding verb is transitive.¹⁰

To understand the predictions of surprisal in this situation, let us refine our grammar very slightly by explicitly distinguishing between transitive and intransitive verbs. We do so by replacing the portion of the grammar of Table 1 that mentions the verb category (V) with a finer-grained variant:

¹⁰Mitchell (1987) and van Gompel and Pickering (2001) originally argued that the differential difficulty effect seen at *the vet* was evidence that transitivity information is initially ignored, but the analysis presented here demonstrates that this effect arises under surprisal when transitivity information is *taken into account*. A related piece of evidence is provided by Staub (2007), who shows that the absence of a comma preceding *the vet* increases processing difficulty; the account here is also generally consistent with this result, since most such subordinate clauses *do* in fact end in commas.

VP → V NP	0.5	Replaced by ⇒	VP → Vtrans NP	0.45
VP → V	0.5		VP → Vtrans	0.05
V → scratched	0.25		VP → Vintrans	0.45
V → removed	0.25		VP → Vintrans NP	0.05
V → arrived	0.5		Vtrans → scratched	0.5
			Vtrans → removed	0.5
			Vintrans → arrived	1

In essence, the revision to the grammar says that verbs come in two varieties: transitive (*scratched* and *removed*) and intransitive (*arrived*); transitive verbs usually have a right-sister NP (but not always); intransitive verbs rarely have a right-sister NP (but not never; e.g., *arrived the night before*). For this revised grammar, surprisals at the ambiguous-NP onset and the disambiguating verb can be found in Table 2. The disambiguating verb is more surprising when the subordinate-clause verb was transitive (*scratched*) than when it was intransitive (*arrived*), reflecting the stronger preceding commitment to the incorrect analysis held in the transitive case. Furthermore, the ambiguous-NP onset is more surprising in the *intransitive* case. This latter effect may be less intuitively obvious: it reflects the fact in the intransitive case, the comprehender must resort to a low-probability grammatical rule to account for the ambiguous-NP onset—either the intransitive verb has an NP right sister or a subordinate clause without a final comma. Hence under surprisal theory the simple act of encoding verb transitivity into a probabilistic grammar accounts for *both* of the processing differentials observed by Staub (2007).

Of course, one may reasonably object that this result obtained by a hand-constructed PCFG might not generalize once a BROAD-COVERAGE grammar with rule probabilities reflecting naturalistic usage is adopted—contemporary probabilistic parsers have rules numbering in the tens of thousands, not in the dozens as in the small PCFGs here (Charniak, 1996). Thus Figure 3 reports region-by-region surprisals alongside the first-pass time results of Staub (2007) using a grammar obtained from the entire parsed Brown corpus (Kučera and Francis, 1967; Marcus et al., 1994) using “vanilla” PCFG estimation (Charniak, 1996; Levy, 2008a). Because the parsed Brown corpus does not mark verb transitivity, I added the single refinement of distinguishing between verbs which do and do not have an immediate right-sister NP; the resulting grammar has 11,984 rules. With such a grammar there is a huge number of incremental parses that are possible for most partial-sentence inputs, so exact analysis is not as simple as for the small grammar of Table 1. Nevertheless, algorithms from computational linguistics can be used to compute word-by-word surprisals for such a large grammar (Jelinek and Lafferty, 1991; Stolcke, 1995). As can be seen in Figure 3, broad-coverage surprisal correctly predicts the two reliable differences in first-pass times: those at the onset of the main-clause subject NP, which does not itself involve garden-path disambiguation; and those at the main-clause verb, which does.¹¹

¹¹This simple model *fails* to capture the empirical result that the garden-path disambiguation effect is larger in magnitude than the surprise effect at the onset of the ambiguous NP. Among other reasons, this failure is due to the fact that specific verb-noun preferences are not encoded in the model. *Arrived* can occasionally have an NP right sister; humans know that *vet* is not a good head for such an NP, but our

Original PCFG		Transitivity-distinguishing PCFG		
Condition	Resolution	Condition	Ambiguity onset	Resolution
NP absent	4.85	Intransitive (arrived)	2.11	3.20
NP present	2.32	Transitive (scratched)	0.44	8.04

Table 2: Surprisals at ambiguity resolution in (16) and (16a), and at ambiguity onset and resolution in (17), using small PCFG

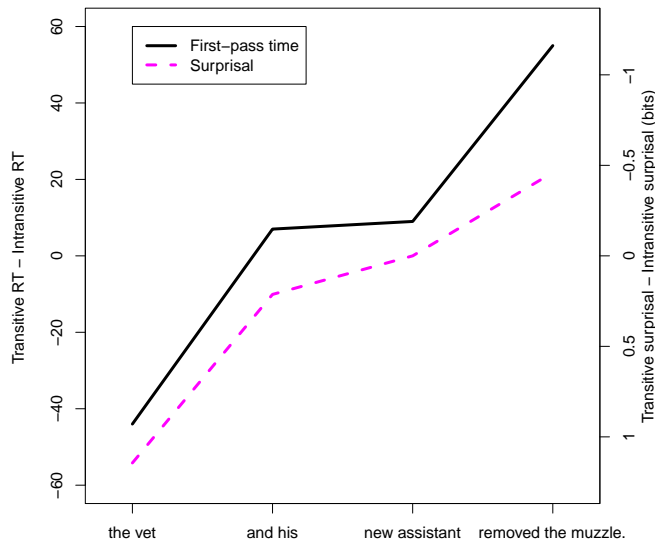


Figure 3: Broad-coverage transitivity-distinguishing PCFG assessed on (17)

Theoretical justifications for surprisal

Another development since Hale’s (2001) original proposal has been work justifying surprisal as a metric for online processing difficulty within the context of rational cognitive models (Shepard, 1987; Anderson, 1990; Tenenbaum and Griffiths, 2001). Here I briefly describe three such justifications in the literature. First, Levy (2008a) posed the view of surprisal as a measure of RERANKING COST. In this approach, the problem of incremental disambiguation is framed as one of allocating limited resources (which correspond to probability mass in probabilistic frameworks) to the possible analyses of the sentence. On this view, the processing difficulty of a word w could be taken to be the *size of the shift* in the resource allocation (equivalently, in the conditional probability distribution over interpretations) induced by w . Levy (2008a) showed that under highly general conditions in which possible joint word-sequence/interpretation structures are specified by a generative probabilistic grammar, the

model does not. Thus our model has not disconfirmed the incorrect analysis before the disambiguating region as fully as a model with finer-grained information sources would have.

size of this shift induced by w , as measured by the RELATIVE ENTROPY (Cover and Thomas, 1991) between the conditional distributions over interpretations before and after seeing w is also the surprisal of w . This reranking-cost interpretation is extremely close to Hale’s original intuition of processing difficulty as being the work done in disconfirming possible structures.

The other two justifications are distinctive in explicitly modeling the comprehender as a RATIONAL agent—that is, one which makes decisions which optimize its expected effectiveness in operating in its environment—and in directly confronting the problem of how much *time* a rational agent would spend processing each word of a sentence in its context. One justification involves a focus on OPTIMAL PERCEPTUAL DISCRIMINATION, and is particularly well-suited to the problem of analyzing motor control in reading. In many theories of reading (Reichle et al., 1998; Engbert et al., 2005; Reilly and Radach, 2006), LEXICAL ACCESS—identifying the word currently attended to, retrieving its representation from memory, and integrating it into the context—is posited to be a key bottleneck in the overall process of reading a sentence. Formalizing this notion of a lexical-access bottleneck turns out to lead naturally to surprisal as an index of incremental processing difficulty. A simple formalization is given for the isolated word-recognition case by Norris (2006, 2009): before encountering any word, the reader has a set of expectations as to what that word may be. What it means to process a word is to accrue noisy perceptual samples from the word; in general, these samples will gradually guide the reader toward correctly recognizing the word. A simple decision rule for the comprehender is to collect input samples until some predetermined threshold of certainty is surpassed, after which the comprehender can confidently commit to the word’s identity and move on. This formulation casts word recognition as a SEQUENTIAL PROBABILITY RATIO TEST, an old problem whose psychological applications date back to Stone (1960; see also Laming, 1968). Mathematical analysis reveals that it is equivalent to a directed random walk in the space of possible word identities, with the average step size toward the correct word identity being approximately constant in log-probability. The starting position of this random walk is simply the word’s surprisal, hence the expected time to decision threshold is linear in log-probability. Figure 4a illustrates example outcomes of this random walk for different surprisal values; note that as word surprisal increases (smaller values on the y axis), smaller changes in raw starting probability are needed to obtain similar changes in the amount of time needed to reach threshold. The only enhancement to this model required to account for sentence-level reading is to make the comprehender’s prior expectations for a word’s identity depend on context, as we saw how to do in the previous section (see Bicknell and Levy, 2010, 2012 for recent work using such a model).

The other rational-analysis justification is of OPTIMAL PREPARATION, introduced by Smith and Levy (2008,). In this approach, one makes the assumption that *some* time-consuming mental computations are required to integrate a word into its context during comprehension; specific commitments as to the type of computation are not required. The time required for a computation is a quantity that can be chosen by the rational agent; it is assumed that shorter times require greater investment of some kind of cognitive resources (these could range from short-term attention to the long-term devotion of specialized neural

pathways), but are also of benefit to the agent in comprehension. This sets up a cost-benefit tradeoff between, on the one hand, investment of resources in the possible inputs that could be encountered in a sentence, and on the other hand, the uncertain payoff obtained from greater processing efficiency on the input that is actually encountered. When this tradeoff is combined with a *scale-free* assumption that the optimal cost-benefit balance is independent of the granularity (e.g., at the level of phrases, words, or syllables) at which investments are made, it results in the prediction that the optimal resource allocation will lead to processing times linear in the log-probability of the event.

CONFLICTING PREDICTIONS BETWEEN EXPECTATIONS AND MEMORY

Because memory- and expectation-based approaches to comprehension difficulty are each supported by deep intuition, theoretical formalization, and a range of empirical results, it is of great interest to examine the degree of overlap in their empirical coverage. In many types of syntactic configurations investigated in the sentence-processing literature, the two approaches make similar predictions regarding differential processing difficulty. As just one example, the pattern of garden-path disambiguation observed by Grodner et al. (2002) in examples like (8) was explained under DLT as a stronger preference to avoid syntactically complex analysis of new input (a reduced relative clause) when memory load is already high (inside a relative clause) than when memory load is lower (inside a sentential complement). This pattern turns out to be predicted under expectation-based disambiguation accounts such as surprisal for two reasons. First, in English RC postmodifiers of embedded-clause subjects (*the valley* in (8)) are simply less common when the embedded clause itself is an RC (the conditional probabilities are less than 0.5% versus 1-2% respectively based on estimates from the Penn Treebank). This consideration leads to the same predictions as do memory-based theories for processing of these structures. Second, the head noun (*valley*) in the RC structure creates a source of expectation for the embedded-clause verb, but in this case the expectation is violated in the RC variant relative to the SC variant (in essence, *captured* is not the verb one expects to have both *river* as its object and *valley* as its subject). Hence *captured* is especially surprising in the ambiguous variant of (8a), which could account for the processing difficulty observed by Grodner et al. (2002).¹² More generally, syntactic configurations which place a heavy load on working memory according to the theories covered in the section on *Memory limitations, locality, and interference* seem to be rare—especially in English (Gildea and Temperley, 2007)—so expectation-based theories predict that they are surprising and thus hard to process.

¹²One weakness of this second account for the data of Grodner et al. is that it predicts a processing difficulty *reversal* further downstream: the greater implausibility of the finite-verb analysis for the RC context should guide the comprehender toward a reduced-relative analysis, which would lighten the processing burden at the disambiguating region *by the enemy*. However, no such processing reversal was found by Grodner et al.. Under surprisal, it is possible that the first consideration (the structural-frequency difference) could eliminate this difficulty reversal. Grodner et al. (2011) present further data bearing on these issues.

Constrained syntactic contexts

But there are also situations where the two approaches can be put into fairly stark conflict. Particular attention has been paid in this regard to SYNTACTICALLY CONSTRAINED CONTEXTS. These are contexts which allow a comprehender to infer that a grammatical event of some type X will occur at some point in future input, but the comprehender is uncertain about exactly when X will occur, and by what surface input (e.g., which word) X will be instantiated. This situation is schematized in Figure 5. Consider a situation with more dependents preceding X (Figure 5b), as compared with a situation with fewer dependents preceding X (Figure 5a). For memory-based theories, processing of X should be more difficult in the case with more dependents, due to the greater number of integrations, greater distance from X of early dependents and/or potential interference among dependents during retrieval. For expectation-based theories, on the other hand, the additional information obtained from more preceding dependents implies that the expectations of the comprehender regarding when X will be encountered and what input will instantiate it will generally be sharper and more accurate; thus there should on average be *less* processing difficulty at X than in the situation with fewer preceding dependents. By looking at processing behavior (e.g., reading times) when the comprehender reaches X , we can hope to gain insight into the relative roles of expectations and memory in online language comprehension.

However, experimental work on such syntactically constrained contexts using different languages and different construction types has not yielded a fully consistent picture: in some cases, the picture looks like that predicted by expectation-based accounts; in other cases, it looks like that predicted by memory-based accounts. Let us begin with some of the clearest evidence of expectation-based processing patterns. In obligatorily head-final language/construction combinations such as the verbal dependency structure of Japanese, Hindi, and German (excepting in the last case main-clause finite verbs), there is little to no evidence that adding preverbal dependents makes processing of the final verb more difficult—rather, these additional dependents seem to make the final verb *easier* to process! To take one example, Konieczny and Döring (2003) examined obligatorily verb-final German subordinate clauses such as in (19) below.

- (19) a. ...dass [der Freund] [dem Kunden] [das Auto] [aus Freude] *verkauft* ...
...that [the friend.NOM] [the client.DAT] [the car.ACC] [of plastic] *bought* ...
“The insight that the friend bought the client the plastic car...”
- b. ...dass [der Freund] [des Kunden] [das Auto] [aus Freude] *verkauft* ...
...that [the friend.NOM] [the client.GEN] [the car.ACC] [of plastic] *bought* ...
“The insight that the friend of the client bought the plastic car...”

In this elegant study, the two variants of the sentence differ only in a single letter, but this character determines whether the second NP of the subordinate clause is dative and thus a dependent of an as-yet-unseen verb (19a) or is genitive and thus a postmodifier of the immediately preceding noun (19b). Regression-path durations on the clause final verb *verkauft* (“bought”) were reliably longer in the genitive-NP variant, suggesting that the dative NP facilitated processing of the verb (see also Konieczny, 1996, 2000 and Levy and

Keller, 2013 for additional related experimental data; and Levy, 2008a for surprisal-based analysis). Similar qualitative effects have been found in Hindi (Vasishth and Lewis, 2006) and Japanese (Nakatani and Gibson, 2008, 2010).

At the same time, there are situations where it seems that it is the predictions of memory-based theories, not expectation-based theories, which are borne out. Postnominal relative clauses with overt relative pronouns are in general syntactically constrained contexts, because the comprehender knows that an RC verb must appear. Levy et al. (ress), for example, parametrically varied the number of intervening constituents between a relative pronoun and the verb in subject-extracted RCs in Russian, such as:

- (20) a. ...ofitsant, kotoryj zabył prinesti bljudo iz teljatiny posetitelju v
 ... waiter, who.NOM forgot to_bring dish.ACC of veal customer.DAT in
 chernom kostjume...
 black suit...
 "... the waiter, who forgot to bring the veal dish to the customer in the black suit..."
- b. ...ofitsant, kotoryj bljudo iz teljatiny zabył prinesti posetitelju v
 ... waiter, who.NOM dish.ACC of veal forgot to_bring customer.DAT in
 chernom kostjume...
 black suit...
- c. ...ofitsant, kotoryj bljudo iz teljatiny posetitelju v chernom kostjume
 ... waiter, who.NOM dish.ACC of veal customer.DAT in black suit
 zabył prinesti...
 forgot to_bring...

Because Russian clause structure has free word order—all logically possible orderings of subject, verb, object, and indirect object are acceptable under some circumstances (Krylova and Khavronina, 1988, *inter alia*)—all three linear orderings in (20) retain the same basic truth-conditional meaning. In (20a), the RC verb complex (*forgot to bring*) immediately follows the relative pronoun; in (20b) the direct object (*the dish of veal*) precedes it; in (20c) both the direct object and the indirect object (*the customer in the black suit*) precede it. Hence expectation-based theories predict that each additional intervener should increase the sharpness of the comprehender's expectations regarding the RC verb complex's argument structure and identity—there are fewer things that a waiter can do to a dish of veal that a waiter can do in general, and even fewer things that a waiter can do to a dish of veal that implicate a customer as an indirect object. Yet there is no trace of expectation-based facilitation at the verb in these cases: instead, reading times at the RC verb complex increase monotonically with the number of intervening constituents. (Levy et al. *did* find expectation-based effects at the processing of the accusative NP, *dish of veal*, which was read more slowly in the RC-initial position, which is a rare position for an accusative NP in Russian RCs, than in the postverbal position, which is a more common position in corpora.) Similar results were obtained for French in a less exhaustive manipulation of RC-internal word order by Holmes and O'Regan (1981).

Finally, let us return to the best-studied syntactically-constrained context of all: English

subject- and object-extracted RCs with a full, definite RC-internal NP:

- (21) a. The reporter that the senator attacked admitted the error.
b. The reporter that attacked the senator admitted the error.

Using a small PCFG, Hale (2001) showed that surprisal predicts greater overall difficulty for the ORC due to the lower frequency in general of ORCs. Under surprisal, however, this greater difficulty should in principle show up as soon as the possibility that the RC is subject-extracted is ruled out—at the onset of the RC subject *the senator*. At the RC verb, in contrast, one would expect the ORC to have the advantage, given that the RC verb’s argument structure and identity are more tightly constrained than for the SRC. The empirical facts in this respect are worth careful attention. It is quite clear from self-paced reading studies that the ORC verb is the site of considerable processing difficulty (Grodner and Gibson, 2005). At the same time, however, Staub (2010) has recently shown that the onset of the ORC is *also* the site of processing difficulty, by comparing sentences like (21) with similar complement-clause sentences such as (22) below:

- (22) The reporter hoped that the senator attacked the chairman of the committee.

In eye-tracking studies, Staub replicated the well-established finding that processing is disrupted at the ORC verb relative to the SRC verb; but at the same time, he also found more regressive eye movements from the very first word of the ORC—the word *the*—than from the same word in the complement clause of (22). Surprisal predicts this effect because a comprehender’s expectation for an NP initiating a complement clause is considerably stronger than that for an NP initiating a relative clause, since the majority of RCs are subject-extracted. Thus we find some suggestion that effects of *both* expectation *and* memory can be observed even in this well-studied construction.

Broad-coverage evaluation of surprisal and DLT

A critical new development over the past several years is BROAD-COVERAGE EVALUATIONS of both expectation-based and memory-based theories by a number of researchers, through analysis of word-by-word reading-time datasets collected using eye-tracking or self-paced reading (Boston et al., 2008; Demberg and Keller, 2008; Frank, 2009; McDonald and Shillcock, 2003; Roark et al., 2009; Smith and Levy, 2008). These broad-coverage evaluations differ from traditional controlled studies in that the materials being read are *complete texts* rather than the isolated sentences typically used in sentence-processing research, they are *naturalistic* (the texts are not constructed for the experiment but are everyday reading materials such as newspaper articles), the potential reading-behavior predictors of theoretical interest are therefore *not balanced*, and the datasets to be analyzed are typically *much larger*, since every word in every sentence has a conditional probability and (according to most linguistic theories) must be integrated into a syntactic representation. These datasets therefore pose special challenges both in quantifying the processing difficulty predicted by a given theory and in analyzing the predictive value of each such quantification. Nevertheless, these

efforts have been consistent in finding significant contributions of surprisal as a predictor in multiple-regression analysis of reading times, even when correlated factors widely known to affect reading behavior such as word length and word frequency (Mitchell, 1984; Rayner, 1998, *inter alia*) are included as controls. Several of these efforts tested specifically syntax-based estimates of surprisal (Boston et al., 2008; Demberg and Keller, 2008); Frank and Bod (2011) compared surprisal estimates based on PCFGs (specifically, the limited-parallel implementation of Roark 2001, 2004) with those based on simple recurrent networks (SRNs; Elman, 1990) and, based on the result that the estimates given by the SRN achieved greater predictive accuracy of word-by-word reading times than those given by the PCFG, suggested that SRNs better describe human sentence-processing performance. Since the estimation of high-quality PCFGs is an open research problem to which a large amount of effort in the field of computational linguistics continues to be devoted, it is clear that an answer to the important question of which models are most psychologically faithful—as assessed by their fit to human reading-time and other comprehension data—is only in its infancy (see Fossum and Levy, 2012; Fernandez Monsalve et al., 2012 for related analyses illustrating that the picture remains incomplete). Demberg and Keller (2008) also constructed a broad-coverage variant of DLT and found that it had predictive value for reading times at nouns and at auxiliary verbs, though curiously not for other words, including open-class verbs. Another broad-coverage analysis, by Smith and Levy (2008,), posed a different question: what is the *shape* of the relationship between conditional word probability and reading time? Surprisal theory assumes a log-linear relationship—that is, reading times should be linear in *log* probabilities, so that a difference between word probabilities of 0.0001 and 0.0099 should have the same effect as that between word probabilities of 0.01 and 0.99. Traditional psycholinguistic practice, on the other hand, implicitly assumes something closer to a linear relationship, with words with in-context Cloze probabilities above 0.6 to 0.7 thought of as “predictable” and those below roughly 0.01 to 0.05 are uniformly categorized as “unpredictable”. In non-parametric multiple regression analyses, Smith and Levy; Smith and Levy recovered a reliable log-linear effect of conditional word probability on reading times in both eye-tracking and self-paced reading datasets, over six orders of magnitude—ranging from probability 1 to 0.000001. In addition to its theoretical value, this result has crucial methodological ramifications, since predictability differences well below 0.01—which could not be reliably recovered from traditional Cloze studies with participants numbering in the dozens—could have large effects on real-time comprehension behavior.

CONCLUSION

This overview of memory and surprisal in human sentence comprehension both sheds light on a wide variety of sentence-processing phenomena and highlights some outstanding questions which require further research. To illustrate what we have learned, turn back to the opening example of the chapter, sentence (1). As I said before, you probably found this sentence most confusing at two places: around the word *admired* and around the phrase *was concerned*. The first source of confusion can be understood as a memory-based integration effect like

those discussed in the section on *Applications beyond center-embedding difficulty*: to process *admired*, you need to simultaneously integrate it with *girl* and *teacher*, neither of which is adjacent to *admired* and both of which share many relevant retrieval cues. The second source of confusion can be understood as an expectation-based surprisal effect like the one we saw in the section on *Surprisal and garden-path disambiguation*: you probably placed your syntactic bets on *her mother* being the object of *call* and thus were surprised to discover, as *was* indicates, that *her mother* was actually the main-clause subject. Thus the two theories we have covered in this chapter resolve the mysteries of why a typical reader finds sentence (1) difficult where she does.

On the other hand, many open questions remain. Why, for example, do we see the discrepancies in incremental processing costs in syntactically constrained contexts across language and construction type described in the section on *Conflicting predictions between expectations and memory*? It is notable that the cases in which the evidence for memory-based processing costs and against expectation-based costs is clearest involves relativization, which as we saw quite early on in this chapter have long been considered to be the basis for canonical examples demonstrating the limitations of human memory capacity in online language comprehension. It is also notable that the clearest evidence for expectation-based patterning in verbal processing comes from obligatorily verb-final languages, in which the comprehender presumably has much more experience with long-distance dependency integrations (see also the comparisons of dependency distances between English and German by Gildea and Temperley, 2010 and Park and Levy, 2009 for more evidence in this connection). Some researchers have taken initial steps toward constructing models which integrate notions of expectation and memory limitation. Demberg and Keller (2009) have introduced an incremental parsing model that contains both *prediction* and *verification* components, which respectively yield surprisal-like and DLT-like processing difficulty gradients. In addition, the model of Lewis and Vasishth (2005) can achieve some types of expectation-derived processing benefits, as processing of multiple preceding dependents can boost the activation level of a governor before it is encountered (see Vasishth and Lewis, 2006 for more discussion).

It is clear that considerably more work—both empirical and theoretical—needs to be done before we have any definitive answers. On the empirical side, coverage of a wider variety of languages and syntactic construction types is required to expand the fundamental knowledge base on which we build theories. On the theoretical side, a number of questions remain outstanding. First, *why* do we see expectation-based patterning in some situations and memory-based patterns in others? What features of the language, construction type, and potentially even comprehension task induce each type of pattern? Second, what features of context are used—and how, and why—to determine a comprehender’s expectations in online comprehension? This has been referred to as the GRAIN-SIZE problem (Mitchell et al., 1995), and while we know that the answer is in general “potentially very fine-grained”, we are still a long way from truly precise answers. Finally: expectation-based models can be understood as the consequence of optimization in comprehension, situating them within frameworks of RATIONAL cognition (Shepard, 1987; Anderson, 1990; Tenenbaum and Griffiths, 2001; see also the section on *Theoretical justifications for surprisal*). To what extent can memory-

based models be understood in a similar light?

ACKNOWLEDGMENTS

I would like to express my gratitude to Ted Gibson, Rick Lewis, and Shravan Vasishth for numerous conversations over the past several years that have improved my understanding of the memory-oriented theories of syntactic complexity described here. All mistakes remain my own, of course; and with luck, I have not misrepresented these theories too badly.

References

- Adams, B. C., C. Clifton, Jr., and D. C. Mitchell (1998). Lexical guidance in sentence processing? *Psychonomic Bulletin & Review* 5(2), 265–270.
- Altmann, G. T. and Y. Kamide (1999). Incremental interpretation at verbs: restricting the domain of subsequent reference. *Cognition* 73(3), 247–264.
- Anderson, J. R. (1990). *The Adaptive Character of Human Thought*. Hillsdale, NJ: Lawrence Erlbaum.
- Arcuri, S. M., S. Rabe-Hesketh, R. G. Morris, and P. K. McGuire (2001). Regional variation of Cloze probabilities for sentence contexts. *Behavior Research Methods, Instruments, & Computers* 33(1), 80–90.
- Attneave, F. (1959). *Applications of Information Theory to Psychology: A summary of basic concepts, methods and results*. New York: Holt, Rinehart and Winston.
- Bever, T. (1970). The cognitive basis for linguistic structures. In J. Hayes (Ed.), *Cognition and the Development of Language*, pp. 279–362. New York: John Wiley & Sons.
- Bicknell, K. and R. Levy (2009, 31 May–5 June). A model of local coherence effects in human sentence processing as consequences of updates from bottom-up prior to posterior beliefs. In *Proceedings of the North American Chapter of the Association for Computational Linguistics–Human Language Technologies (NAACL-HLT) conference*, Boulder, Colorado, USA.
- Bicknell, K. and R. Levy (2010, 11–16 July). A rational model of eye movement control in reading. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, Uppsala, Sweden, pp. 1168–1178.
- Bicknell, K. and R. Levy (2012, 1–4 August). Word predictability and frequency effects in a rational model of reading. In *Proceedings of the 34th Annual Meeting of the Cognitive Science Conference*, Sapporo, Japan.
- Bock, K. and C. A. Miller (1991). Broken agreement. *Cognitive Psychology* 23, 45–93.

- Booth, T. L. (1969, October). Probabilistic representation of formal languages. In *IEEE Conference Record of the 1969 Tenth Annual Symposium on Switching and Automata Theory*, pp. 74–81.
- Boston, M. F., J. T. Hale, R. Kliegl, U. Patil, and S. Vasishth (2008). Parsing costs as predictors of reading difficulty: An evaluation using the Potsdam sentence corpus. *Journal of Eye Movement Research* 2(1), 1–12.
- Charniak, E. (1996). Tree-bank grammars. Technical report, Department of Computer Science, Brown University.
- Chen, E., E. Gibson, and F. Wolf (2005). Online syntactic storage costs in sentence comprehension. *Journal of Memory and Language* 52, 144–169.
- Chomsky, N. (1956). Three models for the description of language. *IRE Transactions on Information Theory* 2(3), 113–124.
- Chomsky, N. (1957). *Syntactic Structures*. The Hague: Mouton.
- Chomsky, N. (1965). *Aspects of the Theory of Syntax*. Cambridge, MA: MIT Press.
- Chomsky, N. (1981). *Lectures on Government and Binding*. Dordrecht: Foris Publishers.
- Chomsky, N. and G. A. Miller (1963). Introduction to the formal analysis of natural languages. See Luce et al. (1963), pp. 269–321.
- Christiansen, M. H. and M. C. MacDonald (2009). A usage-based approach to recursion in sentence processing. *Language Learning* 51(Suppl. 1), 126–161.
- Clifton, Jr., C. (1993). Thematic roles in sentence parsing. *Canadian Journal of Experimental Psychology* 47(2), 224–246.
- Cover, T. and J. Thomas (1991). *Elements of Information Theory*. New York: Wiley.
- Cowper, E. A. (1976). *Constraints on Sentence Complexity: A Model for Syntactic Processing*. Ph. D. thesis, Brown University, Providence, RI.
- Crocker, M. and T. Brants (2000). Wide-coverage probabilistic sentence processing. *Journal of Psycholinguistic Research* 29(6), 647–669.
- DeLong, K. A., T. P. Urbach, and M. Kutas (2005). Probabilistic word pre-activation during language comprehension inferred from electrical brain activity. *Nature Neuroscience* 8, 1117–1121.
- Demberg, V. and F. Keller (2008). Data from eye-tracking corpora as evidence for theories of syntactic processing complexity. *Cognition* 109(2), 193–210.

- Demberg, V. and F. Keller (2009, 29 July–1 August). A computational model of prediction in human parsing: Unifying locality and surprisal effects. In *Proceedings of CogSci*, Amsterdam, Netherlands.
- Eberhard, K. M. (1999). The accessibility of conceptual number to the processes of subject-verb agreement in English. *Journal of Memory and Language* 41, 560–578.
- Eberhard, K. M., J. C. Cutting, and K. Bock (2005). Making syntax of sense: Number agreement in sentence production. *Psychological Review* 112(3), 531–559.
- Ehrlich, S. F. and K. Rayner (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.
- Engbert, R., A. Nuthmann, E. M. Richter, and R. Kliegl (2005). SWIFT: a dynamical model of saccade generation during reading. *Psychological Review* 112(4), 777–813.
- Fernandez Monsalve, I., S. L. Frank, and G. Vigliocco (2012, 23–27 April). Lexical surprisal as a general predictor of reading time. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, Avignon, France.
- Ferreira, F. and J. Henderson (1990). Use of verb information in syntactic parsing: evidence from eye movements and word-by-word self-paced reading. *Journal of Experimental Psychology: Learning, Memory, & Cognition* 16(4), 555–68.
- Ford, M. (1983). A method for obtaining measures of local parsing complexity throughout sentences. *Journal of Verbal Learning and Verbal Behavior* 22, 203–218.
- Fossum, V. and R. Levy (2012, 7 June). Sequential vs. hierarchical syntactic models of human incremental sentence processing. In *Proceedings of the 3rd Annual Workshop on Cognitive Modeling and Computational Linguistics*, Montreal, Quebec.
- Franck, J., G. Vigliocco, and J. Nicol (2002). Subject-verb agreement errors in French and English: The role of syntactic hierarchy. *Language & Cognitive Processes* 17(4), 371–404.
- Frank, S. L. (2009, 29 July–1 August). Surprisal-based comparison between a symbolic and a connectionist model of sentence processing. In *Proceedings of the 31st Annual Conference of the Cognitive Science Society*, Amsterdam, Netherlands, pp. 1139–1144.
- Frank, S. L. and R. Bod (2011). Insensitivity of the human sentence-processing system to hierarchical structure. *Psychological Science* 22(6), 829–834.
- Frazier, L. and K. Rayner (1982). Making and correcting errors during sentence comprehension: Eye movements in the analysis of structurally ambiguous sentences. *Cognitive Psychology* 14, 178–210.

- Gibson, E. (1991). *A computational theory of human linguistic processing: memory limitations and processing breakdown*. Ph. D. 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 A. Marantz, Y. Miyashita, and W. O’Neil (Eds.), *Image, Language, Brain*, pp. 95–126. Cambridge, MA: MIT Press.
- Gibson, E., T. Desmet, D. Grodner, D. Watson, and K. Ko (2005). Reading relative clauses in English. *Language & Cognitive Processes* 16(2), 313–353.
- Gibson, E. and J. Thomas (1997). Processing load judgements in English: Evidence for the Syntactic Prediction Locality Theory of syntactic complexity. Manuscript, MIT, Cambridge, MA.
- Gibson, E. and J. Thomas (1999). The perception of complex ungrammatical sentences as grammatical. *Language & Cognitive Processes* 14(3), 225–248.
- Gildea, D. and D. Temperley (2007, 23–30 June). Optimizing grammars for minimum dependency length. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, Prague, Czech Republic.
- Gildea, D. and D. Temperley (2010). Do grammars minimize dependency length? *Cognitive Science* 34, 286–310.
- Gordon, P. C., R. Hendrick, and M. Johnson (2001). Memory interference during language processing. *Journal of Experimental Psychology: Learning, Memory, & Cognition* 27, 1411–1423.
- Gordon, P. C., R. Hendrick, and M. Johnson (2004). Effects of noun phrase type on sentence complexity. *Journal of Memory and Language* 51(1), 97–114.
- Grodner, D., K. Comer, and E. Gibson (2011). Non-local syntactic influences in structural ambiguity resolution. In preparation.
- Grodner, D. and E. Gibson (2005). Some consequences of the serial nature of linguistic input. *Cognitive Science* 29(2), 261–290.
- Grodner, D., E. Gibson, and S. Tunstall (2002). Syntactic complexity in ambiguity resolution. *Journal of Memory and Language* 46, 267–295.
- Hale, J. (2001, 2–7 June). A probabilistic Earley parser as a psycholinguistic model. In *Proceedings of the Second Meeting of the North American Chapter of the Association for Computational Linguistics*, Pittsburgh, Pennsylvania, pp. 159–166.

- Holmes, V. M. and J. K. O'Regan (1981). Eye fixation patterns during the reading of relative-clause sentences. *Journal of Verbal Learning and Verbal Behavior* 20, 417–430.
- Hsiao, F. and E. Gibson (2003). Processing relative clauses in Chinese. *Cognition* 90(11), 3–27.
- Humboldt, W. (1988/1836). *On Language*. Cambridge: Cambridge University Press. Translated from the German by Peter Heath. Originally published as *Über die Verschiedenheit des Menschlichen Sprachbaues*, 1836, Berlin.
- Jelinek, F. and J. D. Lafferty (1991). Computation of the probability of initial substring generation by stochastic context free grammars. *Computational Linguistics* 17(3), 315–323.
- Jurafsky, D. (1996). A probabilistic model of lexical and syntactic access and disambiguation. *Cognitive Science* 20(2), 137–194.
- Jurafsky, D. and J. H. Martin (2008). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition* (Second ed.). Prentice-Hall.
- Kimball, J. and J. Aissen (1971). I think, you think, he think. *Linguistic Inquiry* 2, 241–246.
- King, J. and M. A. Just (1991). Individual differences in syntactic processing: The role of working memory. *Journal of Memory and Language* 30(5), 580–602.
- Konieczny, L. (1996). *Human sentence processing: a semantics-oriented parsing approach*. Ph. D. thesis, Universität Freiburg.
- Konieczny, L. (2000). Locality and parsing complexity. *Journal of Psycholinguistic Research* 29(6), 627–645.
- Konieczny, L. and P. Döring (2003, 13–17 July). Anticipation of clause-final heads: Evidence from eye-tracking and SRNs. In *Proceedings of the ICCS/ASCS Joint International Conference on Cognitive Science*, Sydney, Australia.
- Krylova, O. and S. Khavronina (1988). *Word Order in Russian*. Moscow, USSR: Russky Yazyk Publishers.
- Kutas, M. and S. A. Hillyard (1980). Reading senseless sentences: Brain potentials reflect semantic incongruity. *Science* 207(4427), 203–205.
- Kutas, M. and S. A. Hillyard (1984). Brain potentials during reading reflect word expectancy and semantic association. *Nature* 307, 161–163.
- Kučera, H. and W. N. Francis (1967). *Computational Analysis of Present-day American English*. Providence, RI: Brown University Press.

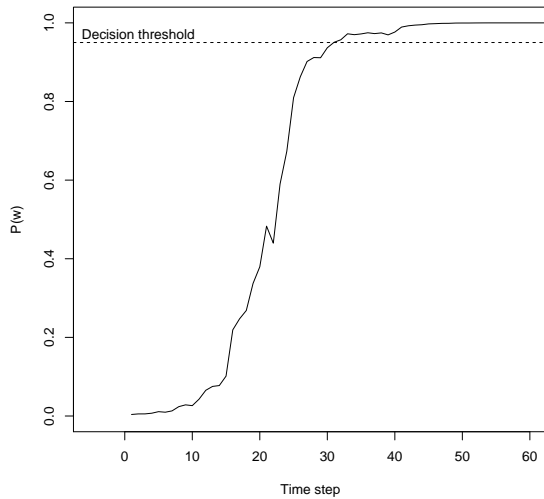
- Laming, D. R. J. (1968). *Information Theory of Choice-Reaction Times*. London: Academic Press.
- Levy, R. (2008a). Expectation-based syntactic comprehension. *Cognition* 106, 1126–1177.
- Levy, R. (2008b, 25–27 October). A noisy-channel model of rational human sentence comprehension under uncertain input. In *Proceedings of the 13th Conference on Empirical Methods in Natural Language Processing*, Waikiki, Honolulu, pp. 234–243.
- Levy, R., K. Bicknell, T. Slattery, and K. Rayner (2009). Eye movement evidence that readers maintain and act on uncertainty about past linguistic input. *Proceedings of the National Academy of Sciences* 106(50), 21086–21090.
- Levy, R., E. Fedorenko, and E. Gibson (In Press). The syntactic complexity of Russian relative clauses. *Journal of Memory and Language*.
- Levy, R. and F. Keller (2013). Expectation and locality effects in German verb-final structures. *Journal of Memory and Language* 68(2), 199–222.
- Lewis, R. L. and S. Vasishth (2005). An activation-based model of sentence processing as skilled memory retrieval. *Cognitive Science* 29, 1–45.
- Lewis, R. L., S. Vasishth, and J. Van Dyke (2006). Computational principles of working memory in sentence comprehension. *Trends in Cognitive Science* 10(10), 447–454.
- Luce, R. D., R. R. Bush, and E. Galanter (Eds.) (1963). *Handbook of Mathematical Psychology*, Volume II. New York: John Wiley & Sons, Inc.
- MacDonald, M. C. (1993). The interaction of lexical and syntactic ambiguity. *Journal of Memory and Language* 32, 692–715.
- Manning, C. D. and H. Schütze (1999). *Foundations of Statistical Natural Language Processing*. Cambridge, MA: MIT Press.
- Marcus, M. P., B. Santorini, and M. A. Marcinkiewicz (1994). Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics* 19(2), 313–330.
- Marks, L. and G. A. Miller (1964). The role of semantic and syntactic constraints in the memorization of English sentences. *Journal of Verbal Learning and Verbal Behavior* 3, 1–5.
- Marslen-Wilson, W. (1975). Sentence perception as an interactive parallel process. *Science* 189(4198), 226–228.
- McDonald, S. A. and R. C. Shillcock (2003). Low-level predictive inference in reading: The influence of transitional probabilities on eye movements. *Vision Research* 43, 1735–1751.

- McRae, K., M. J. Spivey-Knowlton, and M. K. Tanenhaus (1998). Modeling the influence of thematic fit (and other constraints) in on-line sentence comprehension. *Journal of Memory and Language* 38(3), 283–312.
- Miller, G. A. and N. Chomsky (1963). Finitary models of language users. See Luce et al. (1963), pp. 419–491.
- Miller, G. A. and S. Isard (1963). Some perceptual consequences of linguistic rules. *Journal of Verbal Learning and Verbal Behavior* 2, 217–228.
- Mitchell, D. C. (1984). An evaluation of subject-paced reading tasks and other methods for investigating immediate processes in reading. In D. Kieras and M. A. Just (Eds.), *New methods in reading comprehension*. Hillsdale, NJ: Earlbaum.
- Mitchell, D. C. (1987). Lexical guidance in human parsing: Locus and processing characteristics. In M. Coltheart (Ed.), *Attention and Performance XII: The psychology of reading*. London: Erlbaum.
- Mitchell, D. C., F. Cuetos, M. Corley, and M. Brysbaert (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 E. Gibson (2008). Distinguishing theories of syntactic expectation cost in sentence comprehension: Evidence from Japanese. *Linguistics* 46(1), 63–87.
- Nakatani, K. and E. Gibson (2010). An on-line study of Japanese nesting complexity. *Cognitive Science* 34(1), 94–112.
- Narayanan, S. and D. Jurafsky (1998, 1–4 August 2012). Bayesian models of human sentence processing. In *Proceedings of the Twelfth Annual Meeting of the Cognitive Science Society*, Madison, Wisconsin.
- Narayanan, S. and D. Jurafsky (2002, 3–8 December 2001). A Bayesian model predicts human parse preference and reading time in sentence processing. In *Advances in Neural Information Processing Systems*, Volume 14, Vancouver, British Columbia, Canada, pp. 59–65.
- Norris, D. (2006). The Bayesian Reader: Explaining word recognition as an optimal Bayesian decision process. *Psychological Review* 113(2), 327–357.
- Norris, D. (2009). Putting it all together: A unified account of word recognition and reaction-time distributions. *Psychological Review* 116(1), 207–219.
- Park, Y. A. and R. Levy (2009, 31 May–5 June 2009). Minimal-length linearizations for mildly context-sensitive dependency trees. In *Proceedings of the North American Chapter of the Association for Computational Linguistics – Human Language Technologies (NAACL-HLT) conference*, Boulder, Colorado, USA.

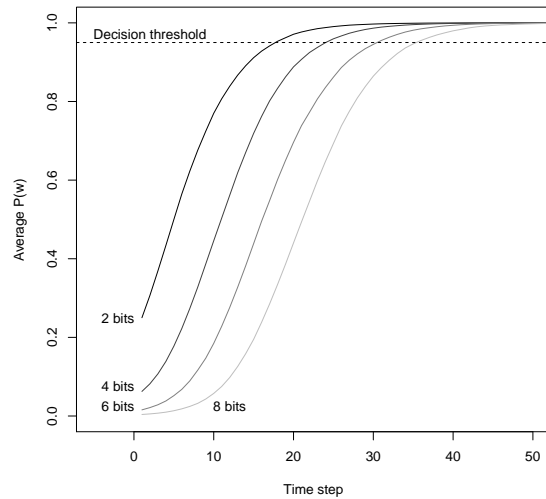
- Pearlmutter, N., S. Garnsey, and K. Bock (1999). Agreement processes in sentence comprehension. *Journal of Memory and Language* 41, 427–456.
- Pickering, M. J. and M. J. Traxler (1998). Plausibility and recovery from garden paths: An eye-tracking study. *Journal of Experimental Psychology: Learning, Memory, & Cognition* 24(4), 940–961.
- Pollard, C. and I. Sag (1994). *Head-Driven Phrase Structure Grammar*. Chicago: University of Chicago Press and Stanford: CSLI Publications.
- Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin* 124(3), 372–422.
- Reichle, E. D., A. Pollatsek, D. L. Fisher, and K. Rayner (1998). Toward a model of eye movement control in reading. *Psychological Review* 105(1), 125–157.
- Reilly, R. G. and R. Radach (2006). Some empirical tests of an interactive activation model of eye movement control in reading. *Cognitive Systems Research* 7, 34–55.
- Roark, B. (2001). Probabilistic top-down parsing and language modeling. *Computational Linguistics* 27(2), 249–276.
- Roark, B. (2004). Robust garden path parsing. *Natural Language Engineering* 10(1), 1–24.
- Roark, B., A. Bachrach, C. Cardenas, and C. Pallier (2009, 6–7 August). Deriving lexical and syntactic expectation-based measures for psycholinguistic modeling via incremental top-down parsing. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, Singapore.
- Shepard, R. N. (1987). Toward a universal law of generalization for psychological science. *Science* 237(4820), 1317–1323.
- Smith, N. J. and R. Levy. The effect of word predictability on reading time is logarithmic. *Cognition*. In press.
- Smith, N. J. and R. Levy (2008, 23–26 July). Optimal processing times in reading: a formal model and empirical investigation. In *Proceedings of the 30th Annual Meeting of the Cognitive Science Society*, Washington, DC.
- Spivey, M. J. and M. K. Tanenhaus (1998). Syntactic ambiguity resolution in discourse: Modeling the effects of referential content and lexical frequency. *Journal of Experimental Psychology: Learning, Memory, & Cognition* 24(6), 1521–1543.
- Staub, A. (2007). The parser doesn’t ignore intransitivity, after all. *Journal of Experimental Psychology: Learning, Memory, & Cognition* 33(3), 550–569.
- Staub, A. (2010). Eye movements and processing difficulty in object relative clauses. *Cognition* 116, 71–86.

- Stolcke, A. (1995). An efficient probabilistic context-free parsing algorithm that computes prefix probabilities. *Computational Linguistics* 21(2), 165–201.
- Stone, M. (1960). Models for choice-reaction times. *Psychometrika* 25(3), 251–260.
- Sturt, P., M. J. Pickering, and M. W. Crocker (1999). Structural change and reanalysis difficulty in language comprehension. *Journal of Memory and Language* 40, 136–150.
- Tabor, W., B. Galantucci, and D. Richardson (2004). Effects of merely local syntactic coherence on sentence processing. *Journal of Memory and Language* 50(4), 355–370.
- Tabor, W. and M. K. Tanenhaus (1999). Dynamical models of sentence processing. *Cognitive Science* 23(4), 491–515.
- Tanenhaus, M. K., M. J. Spivey-Knowlton, K. Eberhard, and J. C. Sedivy (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.
- Tenenbaum, J. B. and T. L. Griffiths (2001). Generalization, similarity, and Bayesian inference. *Behavioral & Brain Sciences* 24, 629–640.
- Traxler, M. J., R. K. Morris, and R. E. Seely (2002). Processing subject and object relative clauses: Evidence from eye movements. *Journal of Memory and Language* 47, 69–90.
- Traxler, M. J., M. J. Pickering, and C. Clifton (1998). Adjunct attachment is not a form of lexical ambiguity resolution. *Journal of Memory and Language* 39, 558–592.
- Trueswell, J. C., M. K. Tanenhaus, and S. M. Garnsey (1994). Semantic influences on parsing: Use of thematic role information in syntactic ambiguity resolution. *Journal of Memory and Language* 33, 285–318.
- Van Berkum, J. J. A., C. M. Brown, P. Zwitserlood, V. Kooijman, and P. Hagoort (2005). Anticipating upcoming words in discourse: Evidence from ERPs and reading times. *Journal of Experimental Psychology: Learning, Memory, & Cognition* 31(3), 443–467.
- Van Dyke, J. A. (2007). Interference effects from grammatically unavailable constituents during sentence processing. *Journal of Experimental Psychology: Learning, Memory, & Cognition* 33(2), 407–430.
- Van Dyke, J. A. and R. L. Lewis (2003). Distinguishing effects of structure and decay on attachment and repair: A retrieval interference theory of recovery from misanalyzed ambiguities. *Journal of Memory and Language* 49(3), 285–316.
- van Gompel, R. P. G. and M. J. Pickering (2001). Lexical guidance in sentence processing: A note on Adams, Clifton, and Mitchell (1998). *Psychonomic Bulletin & Review* 8(4), 851–857.

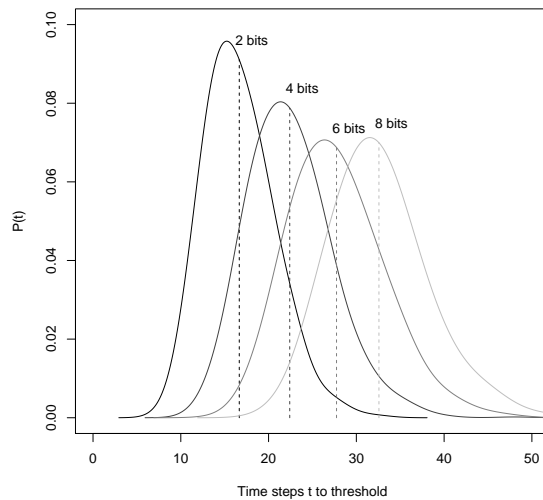
- van Gompel, R. P. G., M. J. Pickering, J. Pearson, and S. P. Liversedge (2005). Evidence against competition during syntactic ambiguity resolution. *Journal of Memory and Language* 52, 284–307.
- van Gompel, R. P. G., M. J. Pickering, and M. J. Traxler (2001). Reanalysis in sentence processing: Evidence against current constraint-based and two-stage models. *Journal of Memory and Language* 45, 225–258.
- Vasishth, S. and R. L. Lewis (2006). Argument-head distance and processing complexity: Explaining both locality and anti-locality effects. *Language* 82(4), 767–794.
- Vasishth, S., K. Suckow, R. L. Lewis, and S. Kern (2010). Short-term forgetting in sentence comprehension: Crosslinguistic evidence from verb-final structures. *Language & Cognitive Processes* 25(4), 533–567.
- Vigliocco, G. and J. Nicol (1998). Separating hierarchical relations and word order in language production: is proximity concord syntactic or linear? *Cognition* 68(1), B13–B29.
- Wagers, M. W., E. F. Lau, and C. Phillips (2009). Agreement attraction in comprehension: Representations and processes. *Journal of Memory and Language* 61, 206–237.
- Wanner, E. and M. Maratsos (1978). An ATN approach to comprehension. In M. Halle, J. Bresnan, and G. A. Miller (Eds.), *Linguistic Theory and Psychological Reality*. Cambridge, MA: MIT Press.
- Warren, T. and E. Gibson (2002). The influence of referential processing on sentence complexity. *Cognition* 85(1), 79–112.
- Wicha, N. Y. Y., E. M. Moreno, and M. Kutas (2004). Anticipating words and their gender: An event-related brain potential study of semantic integration, gender expectancy, and gender agreement in Spanish sentence reading. *Journal of Cognitive Neuroscience* 16(7), 1272–1288.
- Yngve, V. (1960). A model and an hypothesis for language structure. *Proceedings of the American Philosophical Society* 104, 444–466.



(a) Single instance of a random walk

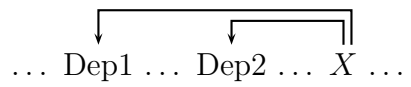


(b) Average posterior probability

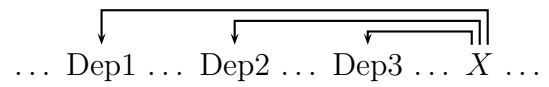


(c) Distribution of times to decision threshold

Figure 4: Surprisal as optimal perceptual discrimination. As time accrues and the word-recognition system accrues more perceptual samples from the current word, the probability of the correct word rises gradually until a decision threshold is reached; changes to raw posterior log-probability accrue more slowly when far from the decision boundary (a). A word’s average posterior probability follows a smooth curve (b), and increase in mean (c, dashed lines) time to recognition is nearly constant in the word’s surprisal. Note that since recognition times are skewed (c, solid lines), mean recognition time is greater than modal time.



(a) Fewer preceding dependents



(b) More preceding dependents

Figure 5: Syntactically constrained contexts with preceding dependents.