LEXICAL STRUCTURE, WEIGHTEDNESS, AND INFORMATION IN SENTENCE PROCESSING

A Thesis

Presented to the Faculty of the Graduate School

of Cornell University

in Partial Fulfillment of the Requirements for the Degree of

Master of Arts

by

Kyle Wade Grove

January 2014

ABSTRACT

This thesis investigates the central question in syntactic theory of how verbal subcategorization is encoded in the lexicon, and how these lexical entries are used during online sentence processing to guide the parse. The thesis proposes the following radical claim: that two types of core sentence processing data (wh-islands, garden path sentences) fall right out of lexical structure differences (+Q,-Q complement selection, and the Unaccusative Hypothesis (Perlmutter, 1978), respectively). On the former, the thesis argues that wh-islandhood is the result of combined ambiguity: the complement-selection ambiguity of +Q *wonder*-type verbs is compounded by the ambiguity of filler-gap sentences. We model this result with by computing the information-theoretic entropy, following (Hale, 2006), of a probabilistic lexical grammar of filler-gap phenomena, with lexical entries for +Q *wonder*-type verbs and -Q *think* type verbs weighted from corpus, and derive greater total entropy reduction in the +Q 'island' condition than in the -Q 'non-island' condition. On the latter, we model an effect from Stevenson and Merlo (1997): reduced relative clauses with embedded unergative verbs (*The horse raced past the barn fell*) are more difficult to process than reduced relative clauses with embedded unaccusative verbs (*The cake baked in the oven fell*). The thesis argues that a co-occurrence relation between the causatives of unergative verbs and prepositional phrases compounds the reduced-relative ambiguity to give rise directly to the garden path effect. We model this result with a probabilistic grammar of reduced relative clauses, with lexical structural entries for unergative and unaccusative verbs weighted from

corpus. We derive greater surprisals (Hale, 2001) for the unergative than the unaccusative case, which we interpret as supporting the account that reduced relative clauses with unergative verbs present the human comprehender with compounded surprise. Overall, the thesis argues that the proper explanatory role of lexical semantics in sentence processing is complexity-based, and not a special appeal to lexically-sensitive ameliorative effects (-Q verbs, as argued in Ross (1967), unaccusative verbs as argued by Stevenson and Merlo (1997)). The thesis argues that the human sentence processor is optimized for human language, and that these classical sentence processing cases are better considered as corner cases where the human sentence processor fails to leverage the lexicon. In such cases, lexical ambiguity compounds main structural ambiguity.

This thesis argues for a resource-optimizing parallel parser where at each point in the sentence, processing resources are allocated proportionally to their probabilistic weight. Information-theoretic accounts of sentence processing (Hale, 2001, 2006; Levy, 2008) obtain on an account of the human sentence processor where resources are dynamically reallocated to more likely and less entropic hypotheses to insure that sentences are parsed as fast as possible with as few resources as possible. The thesis argues for a probabilistic lexical syntax, where the observations of professional lexical semanticists can be encoded in a grammar that can be empirically weighted by corpora and computed by a parser. The thesis strives to maximize parsimony and theory-independence in both the core and linking hypotheses: our theories are formulated as lexicalized mildly context-sensitive grammars, directly suggested by independently attested lexical semantics facts, and translatable into a variety of formalisms (Minimalist Grammars, Tree Adjoining Grammars, Generalized Phrase Structure Grammars); and our results are information-theoretic com-

plexity metrics (entropy, surprisal), which are interpretable as theory-independent measurements, respectively, of ambiguity and surprise.

BIOGRAPHICAL SKETCH

Kyle Wade Grove was born on April 27, 1979 in Bellefontaine, Ohio to Rhonda and Kevin Grove. At the age of seven, he developed skills in his first programming language, *Tandy BASIC*, utilizing a tape recorder to record programs. In middle school, he added to his repertoire by branching out into development in *TI-82 BASIC*. Having mastered BASIC development for Texas Instruments, he took a career break, earning a Bachelor's Degree in English from Wittenberg in 2001 and an MA in Applied Linguistics from Ohio University in 2005. It was at Ohio University that he rediscovered his love of programming, along with a newfound love of cognitive science and linguistic analysis. It was also in 2005 that Kyle applied to the PhD program in Linguistics at Michigan State University, to work on the NSF-funded fMRI Tone Project. Little did he know where that course would ultimately take him...

Kyle Grove now lives with his partner Julie Balazs, Cornell Linguistics 2012, in Roseville, California, where he is Chief Scientist of AI/NLP at AskZiggy, Inc.

ACKNOWLEDGEMENTS

Obviously, the outcome of my graduate process was not the one I originally envisioned. What I ultimately learned is that any ambitious goal worth working towards is also worth reassessing. Where my interests have evolved in a new and fulfilling direction, I have nothing but gratitude, joy, and excitement where there was once angst, frustration and fear.

I owe large debts of gratitude to many people as part of this process. I need to thank my parents Rhonda and Kevin Grove, to whom I owe everything and who gave support for whatever I wanted to do. I need to think my partner Julie Balazs, who guided her own career plans to remain in Ithaca as I worked through graduate school but was only supportive when I decided to leave. I owe my committee, Molly Diesing and Mats Rooth, for diligent attention to my thesis. I am in debt to Michael Putnam, who served as consiglieri to many of my projects. Amongst the many graduate student colleagues from whom I drew daily support and succor: Neil Ashton, David Lutz, Zhong Chen, Marisa Boston, Effi Georgala, Phil Pellino, Greg Johnson, and Elena Cambio. I owe many thanks to the Ohio Program of Intensive English, which provided me with a purposeful sense of career direction early on.

Finally, I owe many thanks to my original adviser, John Hale. While ultimately I came to disagree vehemently with his research program and my place in it, I do have to legitimately say that I would not be the computational linguist I am without him.

TABLE OF CONTENTS

LIST OF FIGURES

Chapter 1

Introduction

Sentence processing is an astonishing task. Berwick and Weinberg (1982) argue that while most linguistic discussion has focused on computationally non-tractable formalisms (Transformational Grammar, GPSG) where worst-case formal parsing complexity is an exponential function of the length of the sentence, sentence processing behavior as measured in the lab is linear; humans parse the bulk of most sentences very rapidly. Humans, in fact, parse most sentences so rapidly that they can usually exhibit predictive behavior of how a sentence will continue given only a part of a sentence, and this fact is exploited in virtually every sentence processing experiment. Thus, most theories of sentence processing revolve around how some cue given in an early part of the sentence is used to winnow down the space of continuation possibilities in the later part of a sentence. Even serial theories of sentence processing have to take the form of using sentence cues to favor one continuation, to be explored, over other continuations, to be ignored.

Miller and Chomsky (1963) posit that the sentence processor contains three interacting modules: a grammar component, a memory component, and a control component. McConville (2001) further elaborates that this grammatical component has syntactic and semantic/world knowledge submodules which can each independently constrain the parse space. It is therefore unsurprising that different camps in modern psycholinguistics each privilege a different module to bear the

explanatory burden of what cues the sentence processor exploits to winnow down the continuation space. It also follows that each camp tends to have a different answer to the other main question in sentence processing: to what extent is the parser parallel?

Experimental psycholinguists in the tradition of Frazier have favored explanations that loosely couple relatively coarse grammars with relatively powerful biases and heuristics constraining the parser to be strictly serial, in what Lewis (2000) describes as deterministic serial models with reanalysis. On this view, the parser's control component must employ powerful biases and heuristics to winnow the continuation space down to a single ongoing parse that the semantic interpreter can handle. For example, Frazier et al. (1983) argued that the parser resolves sentences containing overlapping filler-gap structures by applying what they term 'the Most Recent Filler preference': when in a sentence processing context with multiple extrapositions of syntactic elements, favor discharging the most recent filler at any gap site.

On the other hand, Tanenhaus and colleagues have produced much work ((Tanenhaus et al., 1989; Filip et al., 2001; McConville, 2001)) to suggest that the parser considers and maintains a relatively broad set of mostly equipotential analyses at any one point in a sentence, and that therefore the parser simultaneously leverages a variety of cues, including structural, lexical, functional, and world knowledge cues to deliver a parse. These explanations favor the semantic component's abilities to both interpret the mass of possible parse trees as well as to constrain the parse space. Where the space of parses is constrained by memory and processing

considerations, this tradition is congruent with what Lewis (2000) terms limited parallelism.

In a similar spirit to Tanenhaus and colleagues, a body of work represented by Miller and Chomsky (1963), Pritchett (1992), and Schneider and Phillips (2001) argues that the parser maintains a plural but limited number of analyses at any one moment, but that these analyses are actively prioritized or ranked. This camp, described by Lewis (2000) as ranked limited parallelism, argues that the parser attempts to maintain representations licensed by the grammar, but that the parser is reassessing and re-prioritizing its commitment to the various parses. This group tends to highlight the role of grammatical competence, and not specific heuristics and biases of the parser, as the main locus of explanation in sentence processing. It is with the limited parallelism and ranked limited parallelism camps that this thesis is most closely aligned.

On the fully serial view, the sentence processor has either exclusively focused on a parse, or entirely ruled it out. On the deterministic garden path theory of Frazier and colleagues, the globally correct parse of a sentence can be incorrectly ruled out by an early stage of parsing, such that the parser crashes entirely. Ranked or probabilistic models, on the other hand, permit the possibility of ranking or weighting parses continuously throughout the parsing process. For these ranks or weights to be meaningful in terms of psycholinguistic cost, we must adopt some account of what it means for a parse to be highly weighted. This thesis adopts the view that the parser weights various parses because it seeks to minimize the risk of incremental incomprehension performed; the parser greedily 'wagers' the

most processing resources on the most probable and/or simplest parses. As the parser reranks or reweighs hypotheses, it has to reallocate processing resources to the newly most viable parses, which causes delay and perceived confusion.

As a hypothesis system that can entertain multiple hypotheses, the parser is prone to either false positives or false negatives. The thesis proposes that a parser which is reallocating processing resources on hypotheses has two potential modes of failure: it can commit a false positive by expending too many resources to an ultimately incorrect parse; or it can commit a false negative by spreading resources inefficiently to too large a parse space. The thesis proposes that exactly the first is what happens in the case of surprisal (Hale, 2001; Levy, 2008); on the parallelist view, a garden path effect, as measured by a large surprisal score, occurs when the parser is confronted with much reranking work all at once. This is in some ways analogous to the case of type 1 error: the parser commits resources to a hypothesis that turns out to be false. On the other hand, we propose that when sentences are said to be difficult for working memory reasons, this is exactly analogous to the second failure case: when the parser has to resolve multiple ambiguities where candidate parses are all weighted equiprobably, the result is that every result is equally surprising, because the parser cannot leverage predictive resources towards any one outcome. This latter case is analogous to type 2 error: the parser fails to commit enough resources to a candidate parse which turns out to be ultimately correct.

1.1 Lexicon Structure and the Processor

A reallocating probabilistic parser potentially addresses the conundrum of how exponential-complexity formalisms can account for linear empirically measured processing time; the processor can recruit neural mass to trade off space for time in the fashion described in Berwick and Weinberg (1982), and can optimize on various parameters of the sentence processing problem space. In particular, the lexicon represents one set of facts that can be utilized as explanatory devices for the efficiency of the parser. It is not surprising, then, that parallel theories tend to place more explanatory burden on the lexicon than serialist theories do. A purely serialist view would seem to make the prediction that multiple ambiguities are resolved absolutely independently from each other; structural and lexical ambiguities should never interact. However, because a parallel parser entertains a weighted set of candidate parses, it permits the possibility that prior independent structural and lexical facts can interact to give rise to difficulty in online processing.

1.1.1 Regularity of the Lexicon

The lexicon represents a potential source of optimizations utilized by the human sentence processor to solve an exponential task in sub-linear time. When serialism was at its explanatory zenith, the lexicon was not thought of as a source of possible generalizations; thus the famous quote that the lexicon is above all ‘a prison–it contains only the lawless’ (Di Sciullo and Williams, 1987, 3). This view implies

that rule-based generalization applies only above the level of the lexeme, so it is not clear how the human sentence processor could utilize any element of such a prison of the lawless. However, in hindsight, the lexicon has been found to be much more ruly than anyone would have thought. For the parser to perform such an optimization online, the parser must utilize heuristics stemming from lexical facts. These heuristics would necessarily stem from correlations between lexical facts that the parser can identify and utilize, and their structural and semantic correlates that the parser aims to solve. The enterprise of predicting lexical facts, however, is not the monopoly of any one particular group of linguists. In particular, increasingly sophisticated accounts of lexical regularity have been progressed by three loosely-defined and sometimes conflicting traditions of linguistics: generative and lexical semantics, computational linguistics, and Chomskyan Syntax, particularly Distributed Morphology.

Lexical semanticists have long championed the primacy of semantic roles in mapping deep meaning structures onto phonological surface structures. The 4th century B.C. Indian grammarian Pāṇini developed a grammar of Sankrit that presaged multiple innovations of 20th and 21st century linguistics, including lexical semantics and constraint based grammars. Pāṇini's grammar was a generative grammar in which meanings generated phonological forms as mediated by *karaka*, akin to today's thematic roles (Kiparsky and Staal, 1969) (as cited in (Wechsler, 2006)). Pāṇini's karakas included: *hetu* - Cause; *kartr* - Agent; *apadana* - Source; *karman* - Theme; *karana* - Instrument; *sampradana* - Indirect Object; and *adhikarana* - Locative.

The seminal Vendler (1957) argued for an aspectual classification of verbal events into four types: Activities, States, Achievements and Accomplishments. Vendler distinguished these classes on the bases of *punctuality* and *duration*. Vendler also provided tests for each of these cases, including a forerunner of the 'in an hour'/'for an hour' test for aspect.

Fillmore argued for extending the notion of syntactic case to 'deep case', in which fundamental meaning correlates governed surface subcategorization behavior. In Case Grammar, a particular verb is associated with a particular valence of deep case roles; case roles are hierarchically ordered such that the most prominent present case is promoted to subject position, presaging principles such as Raising (Postal, 1974), Relational Grammar Promotion (Perlmutter and Postal, 1984), and the Extended Projection Principle (Chomsky, 1982).

The Generative Semanticists' program of deep structure through meaning-preserving transformations drew early attention to causative alternation behavior:

(1) a. The desk moved.

b. I moved the desk.

(2) a. John suffocated.

b. I suffocated John.

(Lakoff, 1976, 46)

Lakoff (1976) classified verbs with an inventory of binary semantic features such

as *+/-DS* (do something),*+/-affect*, *+/- effect*, *+/- poss* (possessive), and rules relating these features in an implicational hierarchy. These semantic primitives were motivated by Generative Semantic's primary desideratum of the Paraphrase Principle: sentences which are paraphrases of each other should exhibit identical deep structures and are to be related by meaning-preserving transformations. The Generative Semantics program would culminate in radical lexical decomposition, such that Lakoff (1969) would provide the famous decomposition of the lexeme *kill* as *CAUSE BE NOT alive*.

Perlmutter (1978) made additional arguments for lexical substructure. He observed that intransitive verbs fall into two categories: unergative (accusative) verbs, such as 'run', whose subject is a thematic Agent, patterning after active voice transitive verbs, and unaccusative verbs such as 'bake', whose subject is a thematic Patient or Theme, patterning after passive voice transitive verbs.

In Role and Reference Grammar (Van Valin Jr and Foley, 1980), verbs are lexically decomposed into abstract atomic predicates: CAUSE, BE, BECOME. Apart from these abstract predicates which are inherent components of the verb's core, Role and Reference Grammar does not permit phonetically-null elements such as traces or functional heads. Syntactic analyses in Role and Reference Grammar apart from verb valency therefore are directly realized by morphological and phonological processes, the means of which are not universal but realized distinctly in various languages. The particular language admits licit syntactic forms and operations which realize the verb's 'core' and 'periphery'. Here too, thematic role are hierarchically arranged, and particular languages may tend to front the most

prominent role, but for reasons of discourse prominence rather than reasons of syntactic well-formedness.

Dowty (1991) argued that two central archetypal 'proto-roles' give rise to the plethera of different thematic roles found in the literature, namely, proto-Agent and proto-Theme. Proto-Agent potentially gives rise to Agent, Cause, and Experiencer roles seen in the lexical semantics literature, whereas Proto-Theme gives rise to Theme, Patient, and Goal roles.

Levin and Rappaport-Havov (1995) developed an inventory of verbs classified by syntactic behaviors, including alternation, transitivity, and prepositional phrase attachment. Levin and Rappaport-Havov (1995) were the first to show that unergative verbs potentially behave akin to unaccusative verbs in licensing the transitivity alternation when attached by a Path phrase.

(3) a. ***The window broke.***

b. Pat ***broke the window.***

(4) a. **The soldiers marched** *to their tents.*

b. The general **marched the soldiers*** *(to the tents).* (Levin and Rappaport-Havov, 1995)

Syntacticians in the Chomskyan tradition gradually turned their attention towards lexical matters, particularly as the lexicon and morphosyntax was found to interact with Chomskyan core desiderata: binding and movement. Hoekstra (1988),

extending the Perlmutter (1978) program of unifying morphosyntax and lexical semantics, examined small clause resultatives. Hoekstra (1988)'s Government and Binding Theory analysis of small clauses gave rise to a unifying explanation of intransitive, transitive, and ditransitive verbs, where verb transitivity is not an inherent property of a lexical item but rather an artifact of lexical semantics and syntactic composition.

Larson (1988) examined the syntax of such double complement and double object constructions as the following:

(5) a. John sent the letter to Mary.

b. John sent Mary the letter.

Larson, following Barss and Lasnik (1986), demonstrates binding and scope asymmetries in the double object construction:

(6) a. I showed each man the other's socks.

b. * I showed the other's man each friend. (Larson, 1988, 337)

(7) a. I showed no one anything.

b. * I showed anyone nothing. (Larson, 1988, 337)

Larson observes that if c-command is to be the explanatory mechanism for quantifier scope and negative polarity item binding, then syntactic accounts of the double

object construction and double complement construction where both objects are dominated by the same verbal head are falsified. As shown in 1.1, neither ternary branching structures nor X-bar structures where with one object the specifier and the other the head derive the correct C-command configuration.

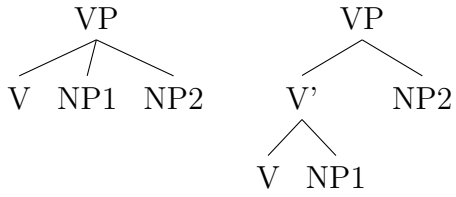

Figure 1.1: Symmetrical Dative Structures of Double Complement Structure

To explain the double complement construction, Larson instead proposes a multiple-head structure where each object is dominated by its own head, and where the surface phonological ordering is obtained by head-movement (V-raising). On Larson's account, the double object construction is a transformation from this base form.

(8) John sent the letter to Mary (Larson, 1988, 342)

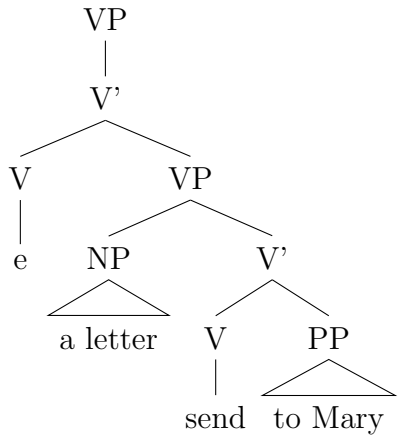

Figure 1.2: Asymmetrical Dative Structure of Double Complement Structure

While lexical decomposition as an approach long predated Larson, their work pioneered the use of syntax in the lexicon as a means of doing lexical semantics. Moreover, the appeal to binding facts and machinery from general syntax in explaining previously 'lexical' facts served to deconstruct the bright line between the syntax and the lexicon. Increasingly, linguists viewed the lexicon not as an explanatory primitive or repository of the lawless, and more as an additional realm where familiar syntactic operations applied and familiar syntactic phenomena were manifest.

Hale and Keyser (1993)'s study of denominal verbs brought the concept of lexical structure (l-syntax) into mainstream syntactic analysis. Hale and Keyser prefigured the Distributed Morphology concepts of roots and verbalizing heads by analyzing intransitive denominals of birthing such as *foaled* as transitive light verbs head-moving a direct object nominal into verb position; transitive denominals such as *shelve* are similarly treated as ditransitives which head-move an argument.

Harley (2002), adapting Pesetsky (1996), adopts a Distributed Morphology framework to provide an alternate explanation of the Larson (1988) facts on double objects and double complement constructions. Harley (2002) criticizes both the Larson (1988) explanation of the double object construction as a transformation of the double complement construction, and the lack of parsimony of the multiple VP. Harley instead proposes that an inventory of functional heads, V_{CAUSE}, V_{BE}, and P_{HAVE} directly compose both the double complement and the double object form. In Distributed Morphology, the derivation composes these functional heads before the surface phonology is derived through the competitive Late Insertion

process. Harley posits that *give* is the Late Insertion form which corresponds to the composition of V_{CAUSE} and P_{HAVE}.

While theoreticians made great advances in the decomposition of verbs, empiricists made great strides as well in developing resources enabling sophisticated studies of nuanced verb behavior in large corpora. Baker et al. (1998) established Framenet, a catalogue/corpus of types of verbs organized by Fillmore's Case Grammar. Palmer et al. (2005) developed Propbank, a section of the Penn Treebank annotated with thematic role tags and organized by verb type. These resources enabled corpus linguists and computational semanticists to build sophisticated probabilistic models of the lexicon. Gildea and Jurafsky (2002) utilized Framenet as a corpus resource for training a classifier which predicted Framenet Case roles with 80% precision. Merlo and Stevenson (2001) trained a system from corpus to classify verbs into clusters on the basis of perceived alternation behavior.

As a result of these theoretical and empirical advances, the lexicon is no longer viewed as a simple repository for the ad-hoc and the stipulative, but as a rule-ordered system whose structure oftchi1999sppen mirrors that of syntactic composition. Where lexical items were once indivisible, 'memorized' units of structure and meaning, with the side effect that they could be sources of special scientific appeal, they now are often derived by a functional morphosyntax or L-syntax.

1.1.2 Lexical Information in Sentence Processing

By adopting the view that syntax and the lexicon operate under fundamentally the same procedures, we allow the possibility that asymmetries in lexical productivity give rise to asymmetries in sentence processing. This thesis thus continues the classic lexicalist tradition in sentence processing, following Fodor (1988); Tanenhaus et al. (1989); Boland et al. (1990); Stowe et al. (1991); Pritchett (1992); MacDonald (1994). This tradition in many ways grew out of a reaction against the syntax-first approach common to the Garden Path Theory, and later, the Recent Filler Heuristic (Frazier et al., 1983; Clifton and Frazier, 1986) explanation for filler-gap processing asymmetries.

Bever (1970) first directed attention towards a type of sentence-processing breakdown which came to be known as the garden-path effect. Bever (1970) noted that disastrous breakdown in the sentence processor can occur when highly preferred initial analyses quickly become incongruent with the globally correct parse. In the case of the reduced relative garden path, a verb which is initially analyzed as an active voice, intransitive verb (*raced*) turns out to be a transitive passivized verb embedded in a misleading relative clause.

(9) The horse raced past the barn.

(10) The horse (which was) raced past the barn fell.

Frazier (1987) argues on the basis of garden-path evidence for a strictly serialist account of the parser, strongly informed by non-lexical, structural biases. On Frazier (1987)'s model, a syntactic module makes strong early commitment on the basis of two structure-only metrics: Late Closure, which attaches incoming sentential material into the current clause; and Minimal Attachment, which minimizes the number of nodes used in the construction of a parse. These heuristics, and only these heuristics, are used to derive an early privileged structure. This means that lexical and pragmatic factors are only considered later in the parse, but can prompt an expensive reanalysis which obviates all previous work on the sentence.

Reduced relative garden path sentences were a chief empirical battleground between lexicalists and the Frazier camp. Pritchett (1992) showed that optionally transitive verbs are more difficult than obligatorily transitive verbs in the reduced relative construction.

(11) a. The horse raced past the barn fell.

b. (Rex knows) the boy hurried out the door slipped.

(12) a. The spaceship destroyed in the battle disintegrated.

b. The bird bought in the store flew away.

c. The children found in the woods were frozen.

Likewise, MacDonald (1994) found an alleviating effect of obligatorily transitive, but not optionally transitive, verbs in the processing of reduced-relative garden path constructions. A reduced relative with an obligatorily transitive verb such as

'captured' in 13a. is easier to comprehend than one with an optionally transitive verb, such as 'fought', in 13b.

(13) a. The ruthless dictator captured in the coup was hated throughout the country.

b. The ruthless dictator fought in the coup was hated throughout the country.

In the obligatorily transitive condition, the active-voice analysis cannot be maintained, because the obligatory transitivity of the verb is incompatible with the active voice, intransitive verb frame. However, in the optionally transitive condition, the sentence processor must build and maintain both the active-voice and reduced-relative analyses to the disambiguation point ('was hated' in 13b.). The effect found in MacDonald (1994) follows on the account of reduced-relative garden paths argued for in this thesis, because the surprise of the reduced-relative reanalysis is only compounded by the surprise of the optionally transitive embedded verb; the obligatorily transitive verb by definition is 'priced in' by definition to the surprise of the reduce-relative structure.

Much of the debate over the role of the lexicon in online sentence processing took place over filler-gap constructions. In filler-gap constructions, the parser must map a non-local dependency between an extraposed element (the 'filler' nominal element) and a position in the subcategorization frame (the 'gap') of a verb to-be-determined. When filler-gap sentences contain complex verb structure such as control verbs, ambiguity can result because there are multiple potential gap sites

the filler could conceivably have been extraposed from. The complexity of this task is demonstrated in 14, where a single prefix (14a. and 14b.) can be disambiguated in two quite different ways.

(14) a. Everyone liked the woman the little child begged to sing those stupid French songs.

b. Everyone liked the woman the little child begged to sing those stupid French songs for.

(Frazier et al., 1983, 203)

(15) a. The child$_1$ begged the woman$_2$ to t$_2$ sing those songs.

b. The child$_1$ begged to t$_1$ sing those songs for the woman.

(Frazier et al., 1983, 203)

With object control (OC) verbs (14a and 15a), the subject of the infinitival 'sing' is the object of the transitive control verb 'begged', but with subject control (SC) verbs (14b and 15b), the subject of the infinitival is the subject of an intransitive control verb. While these conditions can generally be disambiguated at the control verb in non-relativized contexts (15), relativization of the control verb's object removes the disambiguating cue in object control cases. Thus, the prefix 'Everyone loved the woman the child begged to sing those French songs ...' is locally ambiguous, and is disambiguated by either the continuation (... *for.*), in the subject control case, or the sentence abruptly ending (.), in the object control case.

Frazier et al. (1983) argued that the lesser difficulty of SC filler-gap supports an account where a strictly serial parsing strategy is guided by a Recent Filler heuristic. The Recent Filler heuristic holds that in cases where the parser is presented with multiple fillers for a single gap, it expects to map the most recent filler to that gap. When this expectation is defeated by new sentential material, the parser must backtrack until it can recover to an analysis congruent with this material.

Fodor (1988) analyzed the core evidence for the Garden Path/Most Recent Filler theory in Frazier et al. (1983) and Clifton and Frazier (1986). As the Garden Path/Most Recent Filler models posits that the parser has immediate access to only bare structural information, Frazier et al. (1983) and Clifton and Frazier (1986) would predict that in the absence of ambiguity, filler-gap structures should be equally difficult. Yet, Fodor (1988) finds that subjects report unambiguous Distant Filler (DF) sentences to be more difficult than unambiguous Recent Filler (RF) sentences, such as in Example 16.

(16) a. DF This is the woman$_i$ who the child$_j$ tried to speak to GAP.

b. RF This is the woman$_i$ who the child$_j$ forced to speak GAP.

Because the Recent Filler hypothesis, a parallel candidate/serial pursuit theory, does not utilize lexical information in first-pass parsing, it does not readily explain processing asymmetry in unambiguous filler-gap structures which are minimal pairs distinguished primarily by subcategorization. The Recent Filler hypothesis predicts that no reanalysis should occur in either 16a or 16b, and therefore that 16a and 16b should be equally difficult to process.

Boland et al. (1990) provided further evidence against the Recent Filler hypothesis by employing a semantic mismatch paradigm for object control verbs in which the raiser-to-object is potentially an implausible subject of the embedded infinitival. This effect is seen in 17, where 17b is semantically anomalous due only to the infinitival verb; horses and outlaws both make excellent receivers of signals, but horses lack the ability to surrender weapons.

(17) a. The cowboy signaled the outlaw to surrender his weapons quietly.

b. The cowboy signaled the horse to surrender his weapons quietly.

(Boland et al., 1990, 416)

By applying semantic mismatch to filler-gap dependencies, an extraposed element could be manipulated for semantic plausibility as a potential filler for a gap site. Boland et al. (1990) developed Distant Filler sentences which featured such plausibility mismatches, using wh-questions. In these sentences, the recent filler matched the gap-site verb for plausibility, but the distant filler did not.

(18) a. Which outlaw did the cowboy signal to surrender his weapons quietly?

b. Which horse did the cowboy signal to surrender his weapons quietly?

(Boland et al., 1990, 417)

Boland et al. (1990) used an online plausibility monitoring task in which participants were asked to incrementally indicate whether the sentence was plausible. If participants employed the Recent Filler heuristic, then they should be unaware of

plausibility mismatches which obtain only on the Distant Filler structural analysis. Participants detected implausibility immediately, suggesting that they do not rely on a Recent Filler heuristic.

Stowe et al. (1991) found that the Boland et al. (1990) effect is modulated by transitivity bias: transitive-bias verbs generated greater breakdowns in plausibility than intransitive-bias verbs. Subjects were asked to assess plausibility as quickly as possible for filled-gap sentences; with transitive-bias verbs, subjects took more time to assess plausibility and gave more negative answers than in other conditions. The results as a whole strongly suggest a parsing architecture where syntactic structure, lexical structure, and lexical world knowledge are all accessible during first-pass parsing. They also suggest that while syntactic structure, lexical structure, and lexical plausibility all inform which parses the parser pursues, syntactic structure and lexical bias play equally important roles in first-pass generation of candidate parses.

In consideration of this background, this thesis argues that a weighted structured lexicon can explain core phenomenon in verb-focused sentence processing. The thesis argues two classic cases of sentence processing difficulty, namely reduced relative clause garden path effects and wh-islands, are in fact artifacts of lexical ambiguity compounding structural ambiguity. Traditionally, these phenomena have been viewed as intrinsically difficult phenomena which can be ameliorated in special cases. We in fact argue the inverse: unergative reduced relative clauses and filler-gap structures with *wonder*-type verbs are especially difficult combinations of lexical and structural ambiguity; the 'ameliorated' cases are simply structural

ambiguities of the kind the parser solves all the time. By implementing the lexicon as a non-privileged realm of syntactic operations in our grammar, we can directly model how lexical ambiguities can 'feed' classic structural ambiguities.

The plan for the thesis is as follows. Chapter 2 introduces the formal framework for modeling the influence of lexical bias on sentence processing: a probabilistic formal grammar that permits syntactic movement and derivational lexical semantics. Chapter 3 introduces the relevant complexity metrics: surprisal (Hale, 2001; Levy, 2008) and entropy (Hale, 2006), and proposes that surprisal and entropy map onto two largely distinct classes of sentence processing phenomena: 'surprise-type' or type 1 sentence processing difficulties; and 'ambiguity-type' or type 2 sentence processing difficulties. Chapter 4 replicates an experiment from Hale (2003b) to show that Entropy Reduction makes incorrect predictions with respect to garden path sentences, as Entropy Reduction predicts that a garden path sentence and the simple main clause continuation should be equally difficult to process.

In 5 of the thesis, we reexamine the Chomsky and Lasnik (1977) claim that only Q-taking verbs such as *wonder* induce whether-islands, as in 19b.

(19) a. Who did Albert say that they dismissed?

b. # Who did Albert wonder whether they dismissed?

c. ? Which employee did Albert wonder whether they dismissed?

Unlike most approaches to wh-islandhood, the thesis holds weak islandhood to be not a structural fact in and of itself, but as an epiphenomenon of parsing multiple

interacting ambiguites (filler-gap ambiguity, and the continuation ambiguity of +Q *wonder*-type verbs (Bresnan, 1972)). The thesis claims that wh-islands result from the multiple dimensions of ambiguity that the parser is confronted with in +Q but not -Q cases: in +Q cases (islands) the parser must resolve the filler-gap ambiguity together with the lexical ambiguity of *wonder*-type verbs as compounded by the syntactic complexity of the questions embedded under *wonder*, whereas in -Q cases the parser need only resolve a filler-gap ambiguity. We use the notion of entropy (Hale, 2003a) defined over a formal grammar to measure the ambiguity or bandwidth the parser faces in resolving such weak-island sentences.

In Chapter 6 of the thesis, we examine the well-studied case of reduced-relative garden paths, as examined first by Bever (1970).

(20) a. The horse raced past the barn fell. (Bever, 1970)

English verb passivization requires that the verb be underlyingly transitive. However, reduced relative effects are often found with intransitive verbs, including unergatives and unaccusatives. Thus, MacDonald (1994) and Stevenson and Merlo (1997) both observe that often successful recovery from the reduced relative effect requires the parser to reanalyze verb transitivity. Upon encountering the initial verb token *raced*, the parser will assign a main clause, active voice, intransitive analysis to the verb. It is only upon reaching the verb token *fell* that sentence processing difficulty is reported; this is uncontroversially believed to result from the parser having to reassess the initial intransitive, active voice, main-clause parse of the first verb as instead a transitive, passivized, reduced-relative clause verb.

We make the intuitive but original claim that reduced relative garden path cases as described by Bever (1970) are driven by verbal lexical semantics. Stevenson and Merlo (1997) noted that for the reduced relative garden path, the identity of the temporarily ambiguous verb is important:

(21) a. **Unaccusative:** The butter melted in the oven was lumpy.

b. **Unergative:** The horse raced past the barn fell.

Stevenson and Merlo (1997) reported that subjects found reduced relative clause structures containing unergative verbs more difficult to parse than reduced relative clause structures contain unaccusatives. Stevenson and Merlo (1997) identified causation as the locus of explanation for this processing asymmetry. Parsing the reduced relative structure requires the comprehender to segue from an initial intransitive, active-voice parse of the embedded verb to a transitive, passivized verb, they suggested that the asymmetry in processing difficulty results from a production asymmetry in the causative alternation. Stevenson and Merlo (1997) hypothesize, appealing to Hale and Keyser (1998), that causative v applies lexically for unaccusatives, but syntactically for unergatives.

The thesis argues that Stevenson and Merlo (1997) are fundamentally correct, although stipulative, in their explanation. The thesis derives Stevenson and Merlo (1997)'s notion that causation applies in a distinct manner for unaccusatives and unergatives, without special appeal to the lexicon. The explanation offered in this thesis is motivated by independent evidence on causative production asymmetries, first reported in Levin and Rappaport-Havov (1995) and shown below.

(22) a. ***The window broke.***

b. Pat ***broke the window.***

(23) a. **The soldiers marched** *to their tents.*

b. The general **marched the soldiers*** *(to the tents).*

Here, Levin and Rappaport-Havov (1995) show that 'change-of-state' unaccusatives and 'manner-of-motion' unergatives exhibit distinct causative alternation behavior. Unaccusative verbs in the 'change-of-state' class, such as "break", readily participate in causative alternation, with no apparent licensing conditions. On the other hand, unergative verbs in the 'manner-of-motion' class, such as "march", can only be causativized when licensed by a directional prepositional phrase. In general, causative alternation behavior requires Accomplishment Aktionsart, provided inherently by change-of-state Unaccusatives, but in the unergative case provided by the path-PP. The path-PP possesses argument properties and Accomplishment Aktionsart; adjunct attachment does not license the directed motion construction (Zubizaretta and Oh, 2007).

The thesis argues that co-occurence restriction of path-PP phrases on unergative causation requires more disambiguation work to be performed in the unergative case. The prediction of increased processing cost of unergative RRC Stevenson and Merlo (1997) falls directly out from a formal grammar which embodies the Levin and Rappaport-Havov (1995) facts, as measured by an information-theoretic complexity metric.

Chapter 2

Grammars of Movement and the Lexicon

2.1 Minimalist Grammars

The Minimalist Grammars framework of Stabler (1997) formalizes transformational grammars written by Chomskyan linguists for use in psycholinguistic complexity metrics such as the Entropy Reduction Hypothesis. The lexical items which compose Minimalist Grammars are triples of phonetic, syntactic and semantic features. The phonetic feature for any Minimalist lexical item is simply the string yield of that item, and can be null, as in the case of covert functional heads. The operations of Merge and Move manipulate phonetic features and are driven by syntactic features. The features governing Merge are of two kinds, Selected Categories and Selector Features. A lexical item whose leftmost Selector Feature matches the Selected Category of another lexical item Merges to create a derived lexical item whose phonetic feature is the concatenation of those of the children, and whose syntactic head is the Selected Category of the selector item.

Formally, Kobele (2006, 4) gives the following definition of Minimalist Grammars.

V , the alphabet, is a finite set. C at, the set of features, is the union of the following pair of disjoint sets:

$Sel \times Bool$, where for $\star x, 0 \in Sel \times Bool$, we write =x, and call it a selector feature

$\star x, 1 \in Sel \times Bool$, we write x, and call it a selectee feature

$Lic \times Bool$, where for $\star y, 0 \in Lic \times Bool$, we write +y, and call it a licensor feature

$\star y, 1 \in Lic \times Bool$, we write -y, and call it a licensee feature

Lex, the lexicon, is a finite set of pairs v, δ, for $v \in V?$, and $\delta \in C$ at F = (merge, move) is the set of structure building operations

Then,

merge1 $\frac{s::=f\gamma tf,\alpha_1,...,\alpha_k}{st:\gamma,\alpha_1,...,\alpha_k}$

merge2 $\frac{s*=f\gamma,\alpha_1,...,\alpha_k tf\delta,\beta_1,...,\beta_l}{s:\gamma,\alpha_1,...,\alpha_k,ts:\delta,\beta_1,...,\beta_l}$

merge3 $\frac{s:=f\gamma,\alpha_1,...,\alpha_k tf\delta,\beta_1,...,\beta_l}{s:\gamma,\alpha_1,...,\alpha_k,\beta_1,...t:\delta,\beta_l}$

move1 $\frac{s:+f\gamma,\alpha_1,...,\alpha_{i?1},t:?f,\alpha_{i+1},...,\alpha_k}{ts:\gamma,\alpha_1,...,\alpha i?1,\alpha_{i+1},...,\alpha_k}$

move2 $\frac{s:+f\gamma,\alpha_1,...,\alpha_{i?1},t:?f\delta,\alpha_{i+1},...,\alpha_k}{s:\gamma,\alpha_1,...,\alpha_{i?1},t:\delta,\alpha_{i+1},...,\alpha_k}$

The various MG rules can be seen to manipulate string yields of categories which are tuple-valued.

A rule in a formal rewrite grammar generally has two functions: to define how symbols can be licitly combined to yield other symbols, and to define how the string yields of the combined symbols should manifest in the string yield of the combination. In a context-free grammar production such as in the figure below, these functions are overloaded in the → operator.

S → NP VP

Figure 2.1: Context Free Grammar Rewrite Rule

At the symbol combination level, the rule suggests we can combine an NP and a VP to yield an S (equivalently, we can split an S to yield an NP and a VP). At the string yield level, the rule implicitly suggests that the string yield of S should be the string yields of NP and VP, combined in left-to-right order. On becoming aware of this conflation inherent in context-free grammars, one can envision grammatical formalisms, such as in the figure below, in which symbol combination and string yield combination do not run in such a lockstep fashion.

S → NP VP [0,0;1,1;1,0]

Figure 2.2: Multiple Context Free Grammar Rewrite Rule

The above Multiple Context Free Grammar (Nakanishi et al., 1997) rule makes explicit the distinction between "Abstract Syntax", the context-free combination of symbols, and "Concrete Syntax", wherein arbitrary linear rewriting functions can define string yield functions more powerful than simple concatenation. Such productions define operations outside the scope of context free grammars, and for that reason MCFGs are in the Context Sensitive tier of the Chomsky Hierarchy. The Context Sensitive tier contains a vast space of hypotheses; even MCFGs with arbitrary linear rewriting functions, though a strict subclass of Context Sensitive Grammars, are themselves too vast and unrestricted a space to constrain linguistic analysis. Moreover, the parsing time of many possible formalisms in the Context Sensitive tier is polynomial with high-degree, or worse, exponential. These considerations influenced the seminal Mildly Context Sensitive Hypothesis of Avarind K. Joshi and Weir (1992): the right grammatical formalism defining human language has a context-free backbone with some range beyond the context-free grammars, is parseable in polynomial time, and allows for cross-serial dependencies. Mildly Context Sensitive Formalisms allow a symbols location in the abstract syntax tree to be discrepant from its linearization in the resulting string yield, capturing the intuition of 'movement'. Each of the Mildly Context Sensitive Formalisms render this discrepancy between Abstract and Concrete Syntax as a different primitive operation alongside concatenation: Minimalist Grammars (Stabler, 1997) permit a Movement function (alongside Merge) that ranges over categories whose string yield is tuple-valued, allowing 'movement' from one tuple component to another component in the resulting parent; Tree Adjoining Grammar (Joshi, 1987) employs Adjunction of one Tree in another, alongside Combination of trees; and Combinatory Categorial Grammar (Steedman, 1987)

employs type-lifting operators (Combinators) to re-map the parameters of the basic Categorial logic. For the various mildly context sensitive formalisms(Range-1 Mildly context Free Grammar, Context Free Language, Head Grammar, Minimalist Grammar, Tree-Adjoining Grammar, Linear Indexed Grammar, Combinatory Categorical Grammar, and Range-2 Mildly Context Free Grammars), the following containment hierarchy for weak equivalence on the resulting languages emerges from the combined work of (H. Seki and Kasami, 1991; Harkema., 2001; Michaelis., 2001.; A. K. Joshi and Weir.).

$$CFL = \textit{1-MCFL} \subset HL = MHL = TAL = LIL = CCL \subset \textit{2-MCFL}$$

It is an open question whether the same containment hierarchy obtains for strong equivalence, which is especially pertinent to the probabilistic modeler working with mildly context sensitive grammars, who wants to know that models defined in one formalism have portability to another. As for weak equivalence, Guillaumin demonstrated a compiler from Minimalist Grammars to Multiple Context Free Grammars, as shown in the figures below. The compilation example of MG to MCFG makes clear how the Movement operations are homomorphic to Merge operations in Abstract Syntax while employing complex string yield functions over tuple-valued categories, to affect the intuition of 'Movement'. Put another way, each lexical item has a distinct set of syntactic (SYN) features, which uniquely determine a movement chain that the lexical item can participate in, as shown in Hale and Stabler (2005). As a result, the derivation trees of Minimalist Grammars themselves constitute the extension of a context free language, and can be probabilistically weighted as context free grammars can.

Mary::D	t0→ Mary	t0	D	
John::D	t0→ John	t1	=D =D V	
likes::=D =D V	t1→ likes	t2	=V C	
who::D -wh	t2→ ""	t3	D -wh	
::=V C	t3→ who	t4	=V +wh C	
::=V +wh C	t4→ ""	t5	=D V	merge1
	t5→t1 t0 [0,0;1,0]	t6	=D V ; -wh	merge3
	t6→t1 t3 [0,0][1,0]	t7	V ; -wh	merge2
	t7→t6 t0 [1,0;0,0][0,1]	t8	V	merge2
	t8→t5 t0 [1,0;0,0]	t11	+wh C; -wh	merge1
	t11→t4 t7 [0,0;1,0][1,1]	t12	C	move2
	t12→t11 [0,1;0,0]			
	t12→t2 t8 [0,0;1,0]			
	S→t12 [0,0]			

Figure 2.3: Compiling Minimalist Grammars to Multiple Context Free Grammars

2.1.1 Probabilistic Grammars

Suppes (1970) introduces the notion of a probabilistic grammar as a model of language competence as deployed in human sentence processing. Whereas a non-probabilistic grammar is simply a constructive definition of a set of grammatical sentences, a probabilistic grammar can assign probabilities to the members of this set. It is able to do so because each rule is annotated with a probability that it will 'fire'. In a non-probabilistic grammar, the sequence of rules chosen in a derivation constitute a constructive proof of that sentence's grammaticality, whereas a probabilistic grammar the product of rule probabilities provides a probability of the sentence. Thus, a probabilistic grammar can deliver on the intuition that certain sentences are likelier than others.

1.0	S	→ NP VP
1.0	PP	→ IN NP
1.0	RRC	→ Vpart PP
0.50	VP	→ Vpast
0.50	VP	→ Vpart PP
1.00	DT	→ “the”
0.50	NN	→ “horse”
0.50	NN	→ “barn”
0.50	Vpart	→ “groomed”
0.50	Vpart	→ ‘*raced*’
0.50	Vpast	→ ‘*raced*’
0.50	Vpast	→ ‘*fell*’
1.00	IN	→ “past”
0.88	NP	→ DT NN
0.12	NP	→ DT N3
1.0	N3	→ NN Z0
0.88	Z0	→ RRC
0.12	Z0	→ RRC Z0

Figure 2.4: Example PCFG: Bever grammar from Hale (2003b)

On the frequentist interpretation of probability (FISHER et al., 1956), these probabilities are taken to be fixed frequencies of the system, recorded in the corpus. On a Bayesian interpretation of probability (Bayes and Price, 1763; Jaynes and Bretthorst, 2003), these probabilities can be understood as degrees of confidence or belief in the productions. The procedure for assigning rule probabilities described in this paper, and common to generative probabilistic models (Bishop et al., 2006), is congruent with both: productions are weighted in accordance with frequency data, but these probabilities are interpreted to be part of the native speaker’s knowledge of the language.

A probabilistic grammar is capable of not only assigning probabilities to the event that a particular rule is used, but can also assign a probability to the event of a particular sentence. Moreover, a probabilistic grammar can assign a probability

to a sentence prefix. This facet of probabilistic grammars is critical for modelling sentence processing experiments, where we care not only about the acceptability of particular sentences but 'where' in the sentence the parser might incrementally encounter difficulty. Hillel et al. (1960) proved that context free grammars are closed under intersection with finite state automata; Billot and Lang (1989) and Nederhof and Satta (2008) provide constructive proofs for intersectioning, respectively, non-probabilistic and probabilistic context-free grammars with finite state machines. Because finite state machines can model sentence prefixes with the forward space of possible continuations, the probabilistic intersection of the prior probabilistic grammar and a finite state machine representing a particular sentence prefix is itself a probabilistic grammar which represents the comprehender's predicted state of knowledge at a particular point in the sentence.

The initial probabilistic grammar thus constitutes a set of prior beliefs the sentence processor has about sentences in general. Each new word causes the sentence processor to redistribute probability mass from some rules to others, occasionally eliminating (setting the probability to zero) rules from the parse. These shifts in probability mass can be quantified using information theory (Shannon, 1948), as described in 3.

2.2 Weighted Lexical Entries

2.2.1 Distributed Morphology

Distributed Morphology (henceforth, DM) (Halle and Marantz, 1993, 1994) eliminates the lexicon as a realm of special operations. DM holds that the processes that underlie lexical composition are essentially syntactic processes, and that lexical entries are themselves derived. Although DM uses only the Minimalist derivational operations of Merge and Move, the DM ontology of representational units requires motivation. First, DM assumes that lexical items do not inherently possess category, but begin as a prelexical derivational item known as a Root. Roots are the only realm of idiosyncratic meaning in DM; they are essentially the manifestation of a Sausserian (De Saussure, 1916) sound-meaning pair, and possess no other syntactic attributes (although they may possess semantic type (Levinson, 2007)). Roots are denoted with a radical symbol, for example, $\sqrt{BREAK}$.

Second, category is not an inherent property of the root element, but is rather a property of the syntactic environment in which the root is embedded (often represented in the verbal realm as v). While roots possess the sound-meaning pair element of lexical knowledge, verb valency and other morphosyntactic elements of lexical meaning are moderated by an economy of functional heads. As Roots do not inherently possess category, morphosyntactic features must be borne independently by a verbalizing element, v (or a nominalizing element n, etc.). A verbal head, $CAUSE$ 1995sbs,pylkk:anen2000rc forms causatives, and projects a specifier for the external argument (Folli and Harley, 2006).

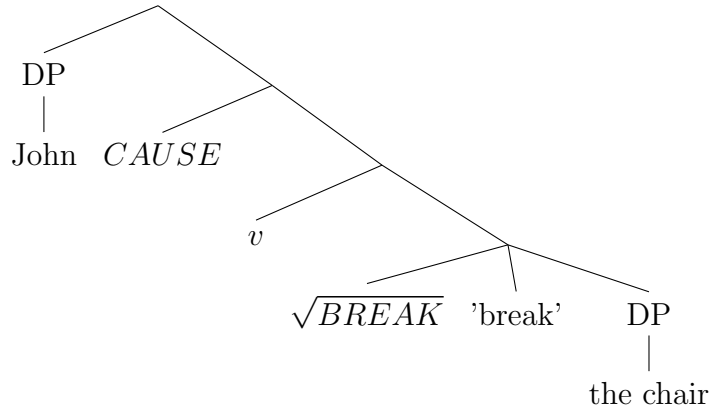

Figure 2.5: DM Causative Analysis

Finally, many subcategorization facts are divorced into a representational component known in the DM literature as the Encyclopedia, a storehouse of lexically-specialized conceptual knowledge. As Distributed Morphology is a sublexical, derivational framework, the inventory of possible words must be limited by some mechanism of the formalism. Distributed Morphology posits the conceptual knowledge of the Encyclopedia as a source of constraint on the range of items and the lexical semantic relations they can enter into in a language. In Distributed Morphology, the Encyclopedia is an (Conceptual-Intentional) (Chomsky, 2000) interface from the derivational morphemic syntax to general world knowledge. As an example, we can imagine unlikely verb forms such as in the following:

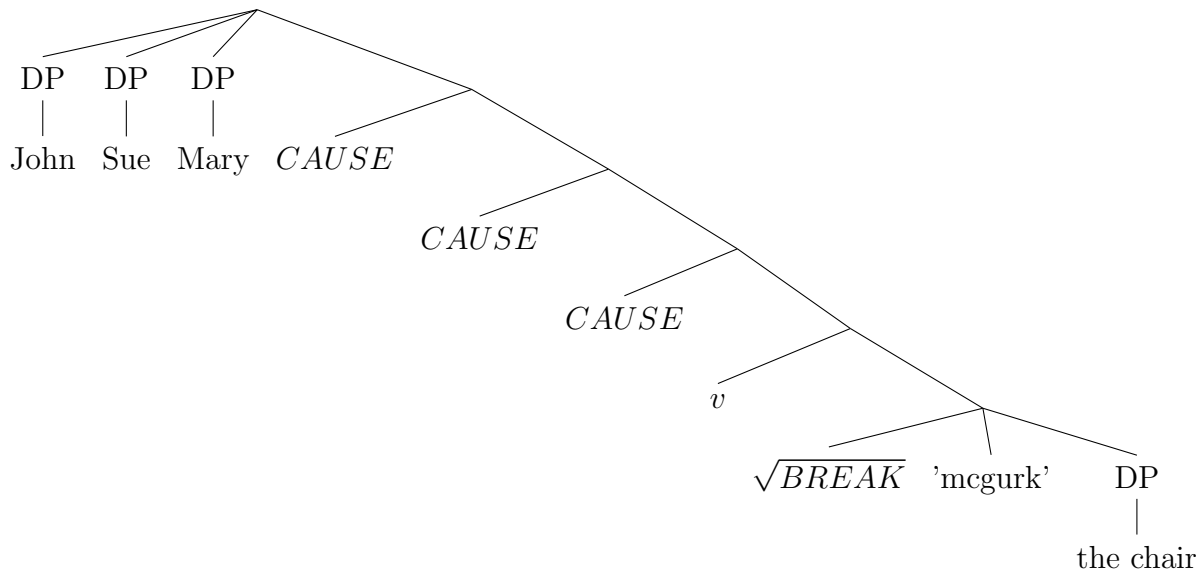

Figure 2.6: Impossible Triple Causative

where a single verb 'mcgurk' is taken to mean 'cause to cause to cause to break', and accordingly assigns four θ roles. On a wholly derivational lexical framework such as Pustejovsky (2005), such a structure would be ruled out by derivational rules banning repeat concatenation of the covert morpheme $CAUSE$. In Distributed Morphology, the derivation of repeat covert concatenation is free to proceed, but is ruled out at the interface by constraints that make such events unlikely and therefore non-economical to depict mono-morphemically. Distributed Morphology can therefore be understood as a generate and filter formalism: generation proceeds derivationally by morpheme concatenation but impossible words are 'blocked' (Embick and Marantz, 2007) by the Encyclopedia.

2.2.2 MG-Distributed Morphology

MG-Distributed Morphology (henceforth, MG-DM) implements the derivational sublexical aspects of Distributed Morphology within the rigorous formal framework of Derivational Minimalist Grammars (Stabler, 1997). Derivational Minimalism assumes lexical items to triples of SYN, SYM, and PHON features, whereas Distributed Morphology explodes these functions of the lexicon to various components of the system: only category-changing heads exhibit syntactic features; Roots implement the concept of the sound-meaning pair but are structurally impoverished; and the Encyclopedia is responsible of grouping together possible words. We implement MG-DM by using MG morphemes with null features to represent this 'distribution' of lexical functions. MG-DM implements Roots as an MG triple with a null (root) SYN feature:

$$\sqrt{run},$$

Figure 2.7: MG-DM Root

To implement covert functional heads such as C, MG permits lexical items to have syntactic and semantic features with zero string yield; this is typically utilized to implement functional heads such as C. MG-DM exploits zero string yield morphemes to implement verbalizing/nominalizing heads:

$$=\text{D } v, \text{ CAUSE}$$

Figure 2.8: MG-DM Verbalizing Head

Thus, in MG-DM, a 'word' is built up derivationally from Roots and functional heads, via MG-Merge.

=D v, CAUSE

Figure 2.9: MG-DM Derived Lexical Item

2.2.3 Weighted MG-Distributed Morphology

MG-DM only implements the generative aspects of Distributed Morphology; by itself, it lacks the filter aspect of the 'generate and filter' Distributed Morphology system. Weighted MG-Distributed Morphology (henceforth, Weighted MG-DM) implements the filtering functions of the Encyclopedia by allowing probabilities to be attached to lexical productions just as they are with syntactic productions. The natural intuition that probabilities can convey world knowledge follows on a Bayesian interpretation of probability, where a probability denotes a degree of belief or knowledge about a random variable. On the Bayesian interpretation of probability, a probability reflects a lack of knowledge of conditioning factors in a system. For instance, we might believe that a coin flip is in fact a deterministic event, and that the perceived randomness of a fair coin is due to our ignorance regarding mediating factors: imperfections in the die that mechanistically steer outcomes towards H or T, but are too myriad and subtle to know. We abstract away from these unknowable factors by using probabilities to stand in for our missing deterministic knowledge of the system as a whole. If we observe that after flipping a coin many thousands of times that the occurrence of H departs

significantly from chance, we might update our belief in the bias of the coin by tilting probabilities towards the perceived outcome, but we remain ignorant of the deterministic conditioning factors affecting that outcome.

With respect to a probabilistic context free grammar, we tend to believe that the grammatical backbone represents the linguist's grammar; it follows then that the probabilities represent non-linguistic facts that bias productions towards one outcome or another, such as is seen with the Encyclopedia. The probabilities in such a grammar represent our non-linguistic 'world knowledge' that one outcome (a possible lexical structure) is more likely to occur than another.

In MG-DM, the 'filter' in the 'generate and filter' mechanism comes from the ability to weight lexical productions from corpora. A supervised training regime utilizing Weighted Relative Frequency Estimation (Chi, 1999) over treebank implements the intuition that the grammar can be 'tuned' through experience with non-linguistic facts.

Chapter 3

Information-Theoretic Complexity Metrics: Surprisal, Entropy, and Total Information

3.1 Surprisal

Following Hale (2001), and Hale (2003a), this thesis employs information theory to model the complexity and strangeness of parser actions. Our core hypothesis for studies in this thesis are Bayesian priors which take the form of probabilistic grammars encoding relevant structural and lexical facts. Our linking hypotheses are information theoretic metrics measuring the severity of updates on this prior as the sentence is encountered incrementally.

Surprisal (Hale, 2001) hypothesizes that perceived difficulty of human sentence processing at a token of interest is associated with the unexpectedness of the new token. On a given string, the surprisal of a token situated between positions *i-1* and i is the logarithm of the ratio of the probabilities of prefixes starting at 0 and ending at *i-1* and i, as shown in the equation below.

$$SURP(w_i) = \log_2 \frac{\alpha_{w_i-1}}{\alpha_{w_i}} \tag{3.1}$$

Surprisal formalizes the intuition that some words are syntactically costly to incorporate, by measuring the rate at which those words reduce the total probability allocated to all incrementally viable analyses. Surprisal predicts garden pathing when new tokens rule out much probability mass.

Surprisal models the cost of syntactically incorporating words into structure, by measuring the rate at which the analysis space shrinks. Because prefix probabilities are defined on probabilistic grammars which themselves constitute statistical processes, surprisal can be interpreted as either an event defined over grammars (Hale, 2001) or an event defined over strings (Levy, 2008). As the prefix probabilities monotonically decrease throughout the sentence, surprisals are positive reals ranging from zero to infinity. A surprisal of zero bits indicated a completely predictable event, such as a string continuation $w+1$ with the same inside probability as w. If w has two equiprobable continuations including $w+1$, the surprisal of $w+1$ is equal to $log2 = 1$ bit.

As a probability measure over parse forests, surprisal generalizes the notion of n-best beam search, which itself generalizes serial processing models such as the Garden Path model (Frazier, 1979). Whereas a Last Resort model pursues the most (1-best) preferred analysis in depth first fashion, surprisal predicts garden pathing when new tokens in the sentence stream disconfirm sections of the beam with great probability mass. On a parallelist view, surprisal models how much

reranking work is triggered by an incoming word, on a n-best ranked beam search generated by a breadth first strategy (Levy, 2008). On a more limited parallelism view, sentences are difficult when a low probability analysis which has fallen off the beam suddenly becomes the only tenable analysis.

3.2 Entropy

Psycholinguists are often interested in the temporary ambiguity found in temporally ordered language stimuli, including garden path sentences. The speech stream is inherently linearly ordered in real time. Psycholinguists exploit this with visual experiments which simulate the speech stream by imposing incremental visual presentation, either in self paced reading or eye-tracking experiments. In the garden path experiment, sentence comprehenders are prompted to abandon a preferred structural analysis because of its incongruity with an incoming word. Garden path sentences thus represent a segue from a multistable perceptual state to a disambiguated perceptual state as the linguistic signal is incrementally encountered.

Ambiguity is a type of uncertainty that language users have about sentences. The linguistic notion of ambiguity, however, does not possess sufficient grain size for use in real-time human sentence processing; we would like a measure of uncertainty with finer grain than the number of trees congruent with the sentence prefix. We would also like to render the notion of preferred and dispreferred analyses in a formal metric. These preferences could be heuristic, but may also be frequency

effects encountered in the acquisition process of the human sentence processor, including recency effects. Stochastic models of sentence processing possess fine grain size, accommodate preference, and are sensitive to experience. Thus, the measure of uncertainty in this paper is the Shannon (1948) entropy, given in the equation below, which operationalizes the notion of ambiguity.

$$\mathrm{h_i} = \mathrm{h}(\epsilon_\mathrm{i}) = -\sum_{\mathrm{r}\epsilon\pi(\epsilon_\mathrm{i})} \mathrm{p_r} \log_2 \mathrm{p_r}$$

Figure 3.1: Entropy of a Discrete Random Variable

Entropy represents the amount of uncertainty about a random variable. For a discrete random variable X with possible outcomes x_1,x_2,..., with probabilities of outcomes p_{x_1},p_{x_2},..., the entropy *H(X)* is equal to $-\sum_{x\in X} p_x \log_2 p_x$. A fair coin, for example, has an entropy of $-((.5\ log_2\ .5) + (.5\ log_2\ .5))$, i.e., 1.0 bits of entropy, since each of the two possible outcomes (Heads, Tails) has a .5 probability of occurring.

As for hierarchical random processes such as probabilistic context free grammars, Grenander (1967) shows how to sum entropy up a product-sum graph using a closed form. For a PCFG with production rules in Chomsky normal form, let the set of production rules in G be Π, and for a given nonterminal ξ denote the set of rules with parent ξ as $\Pi(\xi)$). The entropy associated with a single rewrite of ξ is given by Equation 8.9.

$$H(\xi) = -\sum r \in \Pi(\xi_i) p_r r \log_2 p_r \tag{3.2}$$

A PCFG is a random process whose outcome is a derivation, and the PCFG's total entropy is the entropy associated with derivations in Π, where each derivation is a series of rule selection events. Then the entropy of a PCFG is equal to the total entropy of the start symbol S, where the entropy associated with one-step rewrites of ξ must inherit entropy associated with rewriting children of rules in $\Pi(\xi)$.

Grenander (1967)'s Theorem in Equation 8.10 provides a recurrence relation for determining the entropy of the start category S; each parent accrues entropy from children weighted by the probabilities of those children.

$$H(\xi_i) = h(\xi_i) + \sum r \in \Pi(\xi_i) p_r [H(\xi_{j1}) + H(\xi_{j2}) + ...] \tag{3.3}$$

Granander's Theorem thus gives us a means by which we can compute the entropy of probabilistic grammars via a matrix inversion.

For the purposes of modeling linguistic difficulty, it is helpful to conceptualize three separate aspects of difficulty (ambiguity) which the entropy models. First, entropy models the fact that, *ceteris parabis*, variables with equiprobable outcomes are more difficult than cases where probability is skewed towards certain outcomes. This is seen in Figure 3.2, where the fair coin exhibits greater entropy (1.0 bits)

than a loaded coin with a 0.75 probability of Heads (0.81 bits). The crooked gambler who employs this biased coin with this knowledge has an advantage of 0.19 bits of information over a naive mark who assumes the coin to be equiprobable.

	Biased Coin	Fair Coin
PCFG	0.75 S →  0.25 S → 	0.5 S →  0.5 S → 
H	0.81 bits	1.0 bits

Figure 3.2: Entropy of Variables with Equiprobable vs. Biased Outcomes

Second, entropy models the greater relative difficulty of decisions with more outcomes, when the outcomes are equiprobable. Figure 3.3 depicts a probabilistic grammar modeling a fair coin versus a probabilistic grammar modeling a fair die. Being fair, both the coin and the die have purely equiprobable outcomes, but the die by virtue of having more possible outcomes has greater entropy (2.58 bits).

	Fair Coin	Fair Die
PCFG	0.5 S →  0.5 S → 	0.1 $\bar{6}$ S →  0.1$\bar{6}$ S →  0.1$\bar{6}$ S →  0.1$\bar{6}$ S →  0.1$\bar{6}$ S →  0.1$\bar{6}$ S → 
H	1.0 bits	2.58 bits

Figure 3.3: Entropy of Variables with Few vs. Many Outcomes

Finally, decisions with outcomes dependent on other decisions are more uncertain than simple random variables. Figure 3.4 compares the simple fair coin case to a more convoluted case, where, depending on the outcome of the first coin flip (designated below with start symbol S), the game player must flip (designated below with F) the coin either two or three times. This hierarchical random process has 2.5 bits of entropy, reflecting that the uncertainty of each coin flip is propagating up into the global ambiguity. Hierarchical uncertainty is of particular importance to psycholinguistics because it measures the 'depth' that broadly informs structural metrics in sentence processing.

	Fair Coin	Coin Game
PCFG	0.5 S → 	0.5 S → FFF 0.5 F → 
	0.5 S → 	0.5 S → FF 0.5 F → 
H	1.0 bits	2.5 bits

Figure 3.4: Entropy of Simple Variables versus Hierarchical Random Processes

To find the entropy of prefix grammars, Hale (2006) combines the closed-form computation of PCFG entropy from (Grenander, 1967) with the discovery of Billot and Lang (1989) that intersections of CFGs and automata are themselves CFGs. The probabilistic extension of Billot and Lang (1989) operationalizes the notion of linguistic entropy that a comprehender possesses in an incremental parse of the sentence. That the derivation tree languages of mildly context sensitive grammars are context-free (Hale and Stabler, 2005) allows the extension of this methods to Stabler (1997)'s Minimalist Grammars, as demonstrated in Hale (2006). To operationalize the computation of Probabilistic Minimalist Grammars, Hale (2006)

developed a mini-corpus method in which a small corpus of sentences representing pertinent factors in the experiment was weighted utilizing Weighted Relative Frequency Estimation from measurements on other treebank and corpora and parsed with an MG parser. The resulting parse forest is argued to provide a probabilistic grammar with MG rules weighted by the reference corpora.

The Entropy Reduction Hypothesis (Hale, 2003a) argued that the cognitive load humans experience when comprehending difficult sentences can be modeled as greater decrease in entropy on the incremental structural analysis of the sentence, given some probabilistic grammar known to the sentence comprehender. As the comprehender moves from a greater amount of entropy (H) at one position ($w_i - 1$ in the equation below) to a lesser amount of entropy on the next position (w_i), their perceived effort is argued by Hale (2003a) to increase.

$$ER(w_i) = floor(H(w_i - 1) - H(w_i), 0) \tag{3.4}$$

The incremental difficulty in human sentence processing is modeled by decrease in the conditional entropy of the parse forest conditioned on the string prefix at each point in the sentence. Segues from highly ambiguous states to less ambiguous states mean that he human comprehender has done work, and gained information about the underlying structure of the incoming sentence. A particularly rapid segue of this kind means that the human sentence processor has been particularly taxed by the comprehension task.

3.3 What Surprisal and Entropy Reduction Predict

With the turn in recent years towards probabilistic theories of sentence processing, information theoretical complexity metrics have recieved much attention (Hale, 2001, 2006; Levy, 2008; Boston et al., 2008) in the psycholinguistic literature. While the surprisals and entropies of single random variables are easy to calculate and intuit, not much progress has been made in attempts to flesh out where surprisal and entropy (reduction) make different predictions vis-a-vis sentence processing. Surprisal has been used to model the garden-path effect (Hale, 2001; Levy, 2008), the subject preference effect on non-reduced relative clauses (Levy et al., 2007), and the complexity landscape of standard sentences in eye-tracking corpora (Boston et al., 2008; Bicknell et al., 2009). Entropy or Entropy Reduction has been used to model the difficulty of center embedded sentences (Hale, 2003b), the garden path effect (Hale, 2003b); the Accessibility Hierarchy (Hale, 2006), and the subject preference effect on relative clauses in head-final languages (Yun et al., 2010). Yet, relatively few attempts have been made (though, see Roark et al. (2009) and Wu et al. (2010)) to distill what specific type of sentence processing difficulty these different metrics maps onto.

The advent of probabilistic approaches to the study of sentence processing has coincided with a shift away from an exclusive focus on studying 'laboratory' sentences, towards an increased emphasis on analyzing the normal course of sentence processing on 'ordinary' sentences. This section argues first that the vast majority of sentences which humans parse relatively effortlessly would exhibit small levels of both surprisal and entropy, below some constant threshold which would render

these sentences incomprehensible. Second, and more centrally, the thesis proposes that the 'laboratory' grammatical sentences which humans find difficult fall into two broad classes: 'surprise-type' or type I sentences; and 'ambiguity-type', or type-II sentences, and that each of these subtypes is best modeled, respectively, by surprisal or entropy. Type I sentence processing effects, including garden-path sentences, filled-gap effects, and implausibility effects (Boland et al., 1990), occur when the parser has committed resources to a parse which turns out to be incorrect. Type II sentence processing effects, including center embedded sentences and weak islands, occur when the bandwidth of the parser is exceeded by too many candidate parses, all of which are nearly equiprobable; the parser can not effectively commit predictive resources to the unfolding parse.

This dichotomy falls out of the definition of surprisal as the negative log probability associated with an *outcome* of a random variable, and the definition of entropy as the *uncertainty* associated of a random variable, or the mathematical expectation of surprisals. Surprisals simply reflect the unlikelihood of a certain event, whereas entropies reflect how 'spread out' probability mass is across a slate of the analyses.

More generally, surprisals are therefore maximized for outcomes approaching zero probability, whereas entropies are maximized as the number of possible outcomes increases, and as those outcomes approach equiprobability. Given that human comprehenders typically parse the beginnings of sentences before the ends of sentences, surprisal as a complexity metric is maximized when the early part of a sentence biases the processor to inefficiently allocate processing resources towards

a globally incorrect parse, whereas entropy is maximized when the early part of the sentence is congruent with many possible continuations and yields no real clues as to which of those possible continuations might be the globally correct one.

3.3.1 Entropy Reduction = Work, Surprisal = Waste, Entropy Reduction + Surprisal = Measurable Effort

While the accuracy and speed of human sentence processing is astonishing, study of cognitive processes also reveals biases and inefficiencies. The very concept of a complexity metric assumes an efficient processor where workload is the rate determining step, but a grossly inefficient process can expend much effort to do what amounts to very little work. Psycholinguistic instruments, including acceptability judgments, do not directly measure complexity or work, but rather, the effort expended by a human to comprehend language. The effort we can measure is necessarily a sum of both work performed and effort wasted to perform a cognitive task. Yet very few approaches to modeling human sentence processing provide a calculus for how work and waste contribute to effort.

The central claim of this thesis is that breakdowns in the sentence processor fall into two broad classes measurable by entropy reduction and surprisal, respectively. The constructive aim of the thesis is to provide a calculus where a grammatical fact can predict not only the work complexity of a particular human sentence processing task, but also the effort misallocated on incorrect solutions of the task, and therby a profile of the total effort spent on a human sentence processing task.

Hale (2006) defines Entropy Reduction as a measurement of the amount of work performed in order to disambiguate a sentence. Our account, like Hale (2006), posits Entropy Reduction as a measurement of work, but additionally situates surprisal as a measurement of effort wasted on incorrect hypotheses. We thus posit that the sum of Entropy Reduction and Surprisal, a metric herein termed Total Information, should be a justified and accurate measurement of human sentence processing effort in the general case.

Chapter 4

Empirical Difficulties for Entropy Reduction: the Garden Path Effect

Hale (2003b) offered Entropy Reduction together with the following probabilistic grammar as a possible explanation for the garden path effect reported in Bever (1970).

1.0	S	→ NP VP
1.0	PP	→ IN NP
1.0	RRC	→ Vpart PP
0.50	VP	→ Vpast
0.50	VP	→ Vpart PP
1.00	DT	→ "the"
0.50	NN	→ "horse"
0.50	NN	→ "barn"
0.50	Vpart	→ "groomed"
0.50	Vpart	→ "raced"
0.50	Vpast	→ "raced"
0.50	Vpast	→ "fell"
1.00	IN	→ "past"
0.88	NP	→ DT NN
0.12	NP	→ DT N3
1.0	N3	→ NN Z0
0.88	Z0	→ RRC
0.12	Z0	→ RRC Z0

Figure 4.1: Bever PCFG from Hale (2003b)

Hale (2003b) used a PCFG parser to compute the following profile of Entropy Reduction on the Bever garden-path sentence.

word	*reduction in entropy (bits)*
the	0
horse	1
raced	0.123
past	0
the	0
barn	0.123
fell	3.82

Figure 4.2: Bever ER results from Hale (2003b)

Hale (2003b) argues that the relatively large Entropy Reduction at the token *fell* suggests that the human sentence processor must perform much work to incorporate that token. While the results do indicate that the processor is induced by the garden path continuation to segue from a relatively entropic state to a non-entropic state (completed sentence), the results as presented here confound the disambiguation work associated with the token *fell* with the disambiguation work trigger by the perceived end of the sentence (the wrapping-up effect encountered throughout sentence processing literature (Just and Carpenter, 1980)). This wrapping-up effect can be understood as the difficulty, as measured by Entropy Reduction, encountered by a comprehender when they realize an entropic sentence prefix is to be suddenly analyzed as a finished sentence with entropy zero. This wrapping-up effect obtains on all variety of sentences and is typically an element to be controlled for in sentence processing experiments.

Recall that the garden path effect does not arise from the difficult completion of sentences in general but from the relative difficulty of completing the parse of the

garden path continuation as compared to the main clause continuation. If Entropy Reduction makes the prediction that garden path sentences are as difficult as their main clause continuations, it would fail on any measure to predict the Garden Path Effect. An informal proof, and a replication experiment, follow to demonstrate that this is the case.

Proposition: Entropy Reduction does not predict the garden path effect.

Proof: Let w be a string prefix, ending with an ambiguous verb token, with garden path and main clause continuations of one or more symbols. Let *alpha* be a probabilistic grammar and α_w the probabilistic grammar situated on w. As α_w by assumption is ambiguous and has multiple continuations, then $H(\alpha_w) = \mathrm{C}$ bits where C is non-zero. Let x be an unambiguous main clause continuation, equal to $w + .$ where . is the period or wrapping-up effect, and α_x the grammar situated on x. Likewise, let y be an unambiguous garden-path continuation, equal to $w + z + .$ where z is the disambiguating string or token, and α_y the grammar situated on y. Then, $H(\alpha(x)) = 0$ bits. Likewise, $H(\alpha(y)) = 0$ bits. Let $I((w, x))$ be the Entropy Reduction induced by x on a grammar situated on w, and equal to $H(alpha(w) - H(alpha(x)))$. Then $I(w, x) = H(alpha(w)) - H(alpha(x)) = H(alpha(w)) - 0 = H(alpha(w)$. Likewise, $I(w, y) = H(alpha(w)) - H(alpha(x)) = H(alpha(w)) - 0 = H(alpha(w)$. Then $I(w, x) = I(w, y)$.

Informally, both the main clause continuation with wrapping-up effect and the garden path continuation with wrapping-up effect has the result of disambiguating the ambiguous clause to a string with entropy zero. As the proof above obtains

regardless of prior weightings on $\alpha(x)$ and $\alpha(y)$, it follows that Entropy Reduction effectively ignores the information associated with the garden-path continuation, when the information conveyed is the same in the garden path and main clause continuations. Thus, Entropy Reduction does not predict the Garden Path Effect, as it predicts garden path and main clause continuations to be equally hard.

To reinforce this point, we replicated the Hale (2003b) experiment. We employed the *mcfgcky* parser (Grove, 2010) which uses PCFG renormalization (Nederhof and Satta, 2008) to compute the accurate PCFG situated on the sentence, as described at length in the Appendix. Hale (2003b) utilized a renormalization method in which a category accrued probability iff another category with the same parent were entirely absent in the prefix parse. In this 'naive' renormalization method, a category's probability was reduced to zero iff all paths up to that category were disconfirmed, with its probability shifted to its 'siblings'. *mcfgcky*, employing the (Nederhof and Satta, 2008) inside renormalization algorithm, instead recomputes the entire PCFG whenever any rule is disconfirmed, such that the information propogates through the entire PCFG, manifesting as readjustment (loss) of probability mass for parent categories higher up in the PCFG. For each rule in the PCFG, we compute new probabilities by using the inside algorithm to compute inside probabilities for all categories in the PCFG, multiplying the probability of the original rule by the inside probability of child, and dividing by the product of inside probabilities of the children for that rule, as shown the equations below.

$$P'(A_{(x,y)}) \rightarrow B_{(x_1,y_1)}) = \frac{r(A \rightarrow B)\beta_{B(x_1,y_1)}}{\beta_{A_{(x,y)}}} \tag{4.1}$$

$$P'(A_{(x,y)} \rightarrow B_{(x_1,y_1)}C_{(x_2,y_2)}) = \frac{r(A \rightarrow BC)\beta_{B_{(x_1,y_1)}}\beta_{C_{(x_2,y_2)}}}{\beta_{A_{(x,y)}}} \tag{4.2}$$

We used Hale (2003b)'s Bever PCFG to compute two conditions: the reduced relative clause condition from the original experiment, and a main clause condition which omits the token *fell*. Unlike the Hale parser, for any sentence, *mcfgcky* computes surprisals and entropies for both the prefix ranging over the full sentence and the full sentence itself. The entropy of the former can be non-zero, reflecting the fact that such sentences have possible continuations not realized in the test sentence. Thus, the experimental design separates the entropy associated with the garden path effect from the wrapping up effect, allowing direct comparison to the main clause condition.

word	*reduction in entropy (bits)*
the	0
horse	1
raced	0.123
past	0
the	0
barn	0.123
fell	3.82

Figure 4.3: Hale (2003b): Bever ER Garden Path results

word	*reduction in entropy (bits)*
the	0
horse	0
raced	0
past	0.319
the	0
barn	0
fell	1.494
STOP	0

Figure 4.4: Replication of Hale (2003b): Bever ER Garden Path replication results

word	*reduction in entropy (bits)*
the	0
horse	0
raced	0
past	0.319
the	0
barn	0
STOP	1.494

Figure 4.5: Replication of Hale (2003b): Bever ER Main Clause results

Several facts are apparent. First, the Entropy Reduction encountered in the replication's garden path condition at the disambiguating token *fell* is less than that of the original. Second, more Entropy Reduction is encountered earlier in the sentence. Because PCFG renormalization truly reflects the information conveyed by a word in a sentence by reweighting the entire PCFG whenever a rule is falsified (and not just the associated parent), the replication tends to locate Entropy Reduction earlier in the sentence than the original. This can be thought of as truly measuring the effects of prediction as changes in the uncertainty of the rest of the sentence. Third, the entropy reduction incurred by the disambiguating token *fell* in the garden path condition is exactly that encountered by the sentence end in the main clause continuation. Notably, this equivalence would obtain even where the main clause contained additional sentential material. If the additional material

introduced no additional ambiguity, the entropy reductions would remain equivalent. Even if the additional material increase ambiguity, it would still suggest the spurious prediction that main clause continuations were more difficult than garden path continuations.

Finally, the model is generally restrictive, resulting in the low entropy scores. As a proof of concept, the main facts to be modeled are the entropy of the event space with a reduced relative clause outcome and a main clause outcome. A priori, we have no reason to believe that the results would come out dramatically otherwise if we modeled other facts besides the essential syntax of main clause and reduced relative clause continuations. Where there are other interesting complexities of reduced relative clause continuations, we generally believe they would not effect the main interpretation of the replication here.

Entropy Reduction therefore does not capture the Garden Path effect, as it predicts the main clause continuation and the garden path continuation to be equally difficult. Surprisal here makes the right predictions regarding garden path difficulty. Consider a parse forest with states MC and NP situated over a prefix $0Sw$, and weights α and β on S and NP respectively. We parse the disambiguating token $wVw+1$, which is consistent with NP but not S. Surprisal in this situation predicts minimum difficulty when α is 0, and predicts maximum difficulty (∞) when α is 1. This is in accordance with findings in the psycholinguistics literature in general; the more rare the transition, the more difficult it is to process. However, Entropy Reduction makes an alternative set of predictions. ER is maximized in this case when α and β are each 0.5; it is minimized with either ($\alpha = 1.0$,$beta = 0.0$) or ($\beta =$

1.0,$alpha = 0.0$). Vis-a-vis garden pathing, entropy reduction does not effectively leverage the probabilistic weighting on expectations; as entropy is the expectation of surprisal, it in fact ignores the main clause and garden path continuations by averaging them. Total disambiguation from the weighted parse forest to any single derivation renders the weight allocated on that derivation irrelevant; the entropy of the single derivation is still zero. It is with respect to partial disambiguation, where tokens trigger the reduction of the parse forest to a still-ambiguous parse forest, that Entropy Reduction is in a position to make distinct predictions between outcomes. Such effects can be thought of as the token 'unlocking' a highly ambiguous region of the grammar, giving rise to large Entropy Reduction scores. For the examples that follow, we simply define Cumulative Entropy Reduction as the sum of all one-token Entropy Reductions in the sentence.

0.99	S →	A
0.01	S →	B
0.99	A →	C E
0.01	A →	C F
0.5	B →	D E
0.5	B →	D F
1.0	C →	'a'
1.0	D →	'the'
1.0	E →	'cat'
1.0	F →	'dog'

Figure 4.6: Entropy Increasing PCFG, Initial

The grammar has an associated certainty of 0.171 bits, as Grenander's Theorem yields 0.99 log 0.99 + 0.01 log 0.01 = 0.081 bits for the one-rule rewriting entropy of S, plus 0.99 * ((0.99 log 0.99) + (0.01 log 0.01)) = 0.080 bits for the entropy inherited from A, plus 0.01 * ((0.5 log 0.5) + (0.5 log 0.5)) = .010 bit for the entropy inherited from B. Unlike in sentence-final position, the event of 'a' vs.

'the' matters with respect to Entropy Reduction. The resulting grammar situated on 'a' is rendered in the figure below, with a resulting entropy of 0.081 bits.

1.0	S →	A
0.99	A →	C E
0.01	A →	C F
1.0	C →	'a'
1.0	E →	'cat'
1.0	F →	'dog'

Figure 4.7: Entropy Increasing PCFG, Situated on 'a'

When the event 'the' is seen instead, as seen in 4.8, the grammar conditioned on 'the' has 'unlocked' a more entropic region of the grammar, yielding a situated grammar with greater entropy (1.0) bits than the original, thus yielding a greater possible cumulative entropy reduction.

1.0	S →	B
0.5	B →	D E
0.5	B →	D F
1.0	C →	'a'
1.0	D → '	the'
1.0	E →	'cat'
1.0	F →	'dog'

Figure 4.8: Entropy Increasing PCFG, Situated on 'the'

4.1 Fleshing Out Entropy and Surprisal Predictions

We argue for the classification of sentence processing phenomena into two broad classes of phenomena, to be chiefly explained by surprisal or entropy. Following the

terminology from statistical signal processing, we describe surprisal-type sentence processing effects as 'type-1' effects; the parser has committed a false positive in committing resources to an globally incorrect parse. Likewise, we term entropy-type sentence processing effects as 'type-2' effects; the parser has committed a false negative by not committing resources to any one parse. Type I sentence processing effects, including garden-path sentences, filled-gap effects, and implausibility effects (Boland et al., 1990), occur when the parser has committed resources to a parse which turns out to be incorrect. Type II sentence processing effects, including center embedded sentences and weak islands, occur when the bandwidth of the parser is exceeded by too many candidate parses, all of which are nearly equiprobable; the parser can not effectively commit predictive resources to the unfolding parse.

While sentence-final events with common prefix effectively yield identical Entropy Reductions, sentence-initial and sentence-medial events can give rise to quite different Entropy Reduction profiles cumulative through the remainder of the sentence via the mechanisms described above. We take as granted that the sentence processor is highly optimized for the type of sentences in natural human experience, and that the 'oddball' sentences reported in sentence processing experiments represent deviations from the expectancies such an optimized parser has about sentences. We argue for a realization of this parser as an empirical parser trained on normal sentences; a parallel, predictive probabilistic sentence processor optimized for the likely, simple sentence events that constitute mundane human linguistic experience. We expect not only that sentences found difficult in psycholinguistic experiments should yield unusual entropy or surprisal scores, but that these metrics should pro-

vide useful diagnostics as to how sentence processing failed in any particular case. Where the standard experimental paradigm in sentence processing pits a typical sentence against an 'oddball sentence' in the exploration of exactly one parameter, we expect these oddball sentences to classify as 'surprise'-type sentences or 'ambiguity'-type sentences according to the location and quality of this parameter. We expect 'surprise'-type effects, with surprisal the governing complexity metric, whenever: there exist multiple possible derivations, but one can be said to be favored (on a probabilistic parser, by a higher assigned probability); the parameter is in phrase-final or sentence-final position, where competing derivations can be distinguished by one event; the sentence is initially acceptable until it becomes implausible, and can thus be said to have violated some expectancy. We expect 'ambiguity'-type effects, with Entropy Reduction the governing complexity metric, whenever: there exists multiple possible derivations, but competition between the analyses is said to be a possible factor (as equiprobable analyses constitute a more entropic, ergo, longer to resolve analysis space); the 'top-level' analyses themselves contain subanalyses or interesting recursion that features hierarchical ambiguity; the parameter is early or medial in the sentence; the sentence is from the onset difficult, and can be thought to have exhausted grammatical resources. The above examples suggest that where this oddball parameter is early in the sentence and potentially associated with working memory, we expect Entropy Reduction to play an explanatory role.

4.1.1 Methodological Difficulties: Estimating Entropy Reduction

A consequence for entropy reduction is that the composition of the event macrostate matters in a way that it does not for surprisal. Consider the case where we have a random number generator, X and a possible outcome $X_{=}\ 1$, with an unspecified number of other (mutually exclusive) possible outcomes. As long as the outcomes are mutually exclusive, to compute the surprisal of $(X = 1)$ we need only consider its prior probability, whereas to compute the entropy reduction of $(X = 1)$, we need to know the number of other outcomes and the weights on those other outcomes, which presents comparative difficulties for entropy reduction modeling, particularly with the minicorpus method used in Hale (2003b) and much Entropy Reduction work.

1.0	X →	A B
0.5	A →	C D
0.5	B →	E F

Figure 4.9: Simple PCFG

Imagine the above PCFG as a simplified model of sentence processing, estimated from some corpus. We can derive the surprisal and entropy of a sentence processing event straightforwardly from such a model, but the corpus we estimated from is limited. If we use the mini-corpus method, this is even more worrisome: the experimenter may not have conceived of all possible parameters in the experiment space, and all we have at our disposal for estimating the error in our model are: 1)

traditional inferential statistics such as chi-square over the relative frequencies of independent variables to the model; 2) information-theoretic measurements such as surprisal, relative entropy and cross-entropy which give us information about the validity of the model with respect to the true model.

In the case of surprisal, we can take a surprisal computed from the model and scale it to the surprisal predicted by the empirical distribution by simply multiplying (adding in log-space) the surprisal of the model itself with the surprisal of the corpus given the empirical distribution, as long as all parameters in the model possess accurate relative frequencies (as guaranteed by a chi-square, for a hand weighted grammar, or by the training itself). On the simple PCFG above, there is a certain surprisal indicated on ever symbol for any outcome; estimating the true surprisal of the event can be done by adding the surprisal of the model's start symbol on the event with the surprisal of the model's start symbol in the empirical distribution. Put simply, we scale the surprisal given from the model's microstate by its probability in the empirical macrostate.

We can do no such thing in the case of entropy. The entropy of a microstate given by a hand-weighted grammar has no bearing on the entropy of the macrostate from the empirical distribution, as entropy is a probabilistically weighted-sum. If the hand-weighted grammar misses some rules corresponding to a given parent in the grammar, the unknown probabilities of those missing rules, plus the probabilistic microstate of the missing parameters can affect the error of the resulting entropy computation greatly.

4.2 Total Information

In this section, we define an information-theoretic complexity metric, Total Information, which is simply the sum of Entropy Reduction and Surprisal. This metric combines the Hale (2006) insight that Entropy Reduction represents the amount of information gained segueing from one prefix to the next with the Levy (2008) that Surprisal represents the Kullback-Leibler Distance between one prefix and the next. While Entropy Reduction is a well-founded notion of the amount of *work*, in the physics sense, needed to integrate a new token with the current parse forest, online measures and grammaticality judgment are likely a function of the amount of *effort* needed to integrate a token with the parse forest. We posit that if Entropy Reduction is measuring the amount of direct progress towards the goal, Surprisal represents the amount of wasted effort spent on achieving that goal. On that basis, we submit the metric Total Information as a measurement we believe of the total effort, work or waste, spent in incorporating a token.

$$TI(w_i) = ER(w_i) + SURP(w_i) = floor(H(w_i - 1) - H(w_i)), 0) + \log_2 \frac{\alpha_{w_i-1}}{\alpha_{w_i}} \quad (4.3)$$

That either Entropy Reduction or Surprisal might be the cause of sentence processing difficulty in a particular sentence reflects the intuition that the parser is more efficient in some contexts than in others; a sentence where Entropy Reduction is the chief explanation for processing difficulty is one in which the amount of

work exceeds the bandwidth of even the most efficient parser, whereas a sentence in which Surprisal is the chief explanation of processing difficulty is one in which the 'human' parser is particularly ill-optimized for the given sentence.

Chapter 5

Entropy and Working Memory

5.1 Background

This chapter proposes that entropy-based measures effectively capture processing costs on sentences stated to be difficult for reasons of working memory. In particular, we reexamine the Ross (1967) claim that while interrogative (+Q) CPs are islands for *wh*-extraction declarative (-Q) CPs are not. [1]

Alexopoulou and Keller (2007) report, using a Magnitude Estimation (Cowart, 1997) protocol, that English-speaking subjects find extractions in the Island condition to have degraded acceptability as compared to the Non-island condition, in non-embedded and embedded contexts, as seen in 25.

(24) a. ISLAND ? Who did Mary wonder whether we will fire?

b. ISLAND ? Who did Jane think that Mary wonders whether we will fire?

(25) a. NON-ISLAND Who did Mary claim that we will fire?

b. NON-ISLAND Who did Jane think that Mary claim that we will fire?

Alexopoulou and Keller (2007)'s results quantify what has been known in the classical syntax literature going back to Chomsky and Lasnik (1977); grammatical islands for wh-extraction are sensitive to the nature of the embedded verb. Bresnan (1970) noted two natural classes of sentential-embedding verbs: verbs like *claim*, which subcategorize for -Q complements (*that* or Ø),; and verbs like *wonder*, which subcategorize for a wide array of +Q continuations (*whether*, *if*, and embedded questions), as shown in Example 27.

(26) a. Mary claimed that the supervisor will fire the employee.

b. * Mary claimed whether the supervisor will fire the employee.

c. * Mary claimed if the supervisor will fire the employee.

d. * Mary claimed who the supervisor will fire.

e. * Mary claimed who will fire the employee.

f. * Mary claimed when the supervisor will fire the employee.

(27) a. * Mary wondered that the supervisor will fire the employee.

b. Mary wondered whether the supervisor will fire the employee.

c. Mary wondered if the supervisor will fire the employee.

d. Mary wondered who the supervisor will fire.

e. Mary wondered who will fire the employee.

f. Mary wondered when the supervisor will fire the employee.

Claim-type embedding verbs, which have declarative meaning and take only -Q complements (*that* or) generally allow extraction of embedded *wh*-elements, as seen in 28a. Chomsky and Lasnik (1977) noted that *wonder*-type embedding

verbs, which allow +Q interrogative complements such as *whether* and embedded questions, induce islands for *wh*-extraction (so-called whether islands), as seen in 27.

(28) a. Who did Albert say that they dismissed?

b. # Who did Albert wonder whether they dismissed?

c. ? Which employee did Albert wonder whether they dismissed?

We argue that the Ambiguity Hypothesis accounts for the finding of Alexopoulou and Keller (2007) that *wh*-extraction from +Q embedded clauses produces degraded Magntiude Estimates when compared to *wh*-extraction from -Q embedded clauses. The account is also congruent with several other findings in the literature, which we report in the Discussion section.

Hypothesis: Extraction from embedded interrogatives is difficult because +Q Ambiguity (Bresnan, 1970) compounds the filler-gap ambiguity.

Following Bresnan (1970), we observe that verbs which embed declaratives have predictable complements (Declarative Complement Phrases headed by *that* or ∅), whereas verbs which embed interrogatives subcategorize for a wide array of +Q continuations (Interrogative Complement Phrases headed *whether*, *if*, and embedded questions), as seen in 40. Unlike Chomsky and Lasnik (1977), we claim that it is not necessary to enrich the verbal lexicon with weak-island restrictions, and that the unacceptability of wh-islandhood is a psycholinguistic fact resulting from the failure of the parser to process the increased ambiguity of +Q continuations. We

therefore hypothesize that the incremental ambiguity of *wonder*-class verbs which embed +Q elements results in greater cognitive load for a predictive, incremental parser, giving rise to weak islandhood.

We model results from Alexopoulou and Keller (2007), who found Magnitude Estimates for *wh*-extraction from sententials embedded by *wonder* to be degraded compared to controls with the embedding verb *claim*. We construct and weight a Derivational Minimalist Grammar (Stabler, 1997) which renders the Bresnan (1970) facts. We model the ideal comprehender's performance on the Alexopoulou and Keller (2007) sentences using Entropy Reduction (Hale, 2006). With respect to the modeling factors, sentences in the Island condition convey an average total of 3.379 bits of information, compared to Non-island sentences, which convey 1.899 bits of information [2]

5.1.1 Complement Selection

Selection

Grimshaw (1979)'s *Autonomy Hypothesis* proposed a bifurcated system of verbal subcategorization: c-selection, in which a category selects the syntactic features of

[2]The quantities of entropy and surprisal reported in multiple portions of this thesis may strike the informed reader as small compared to surprisals and entropies derived from broad-coverage grammars. There are two potential causes for these effects, each to be taken into consideration. First, the models are concise with relatively few independent variables. Second, the verbs in the grammars are head-lexicalized, but lexical bias of the selectee of a verb is not generally modeled.

a selectee; and s-selection, in which a category places semantic requirements on its selectee. The necessity of this bifurcation is shown in examples such as 29, where the verb *asked* selects for a +Q complement, but can c-select for either a CP or an NP.

(29) a. John asked me [CP what the time was].

b. John asked me [DP the time].

c. John wondered [CP what the time was].

d. *John wondered [DP the time].

(Adger and Quer, 2001, 108)

Pesetsky (1982) argued that Case can supplant c-selection when coupled with s-selection in a theory of selection. In particular, Pesetsky (1982) explains the existence of verbs which take as complements only concealed question noun phrases and not clausal arguments.

(30) a. It was proved [CP that tomatoes are fruits].

b. *It was proved [NP a theorem].

(31) a. John is curious (about) [IP where I went].

b. John is curious *(about) [NP life].

Clause Embedding Verbs

Bresnan (1970) prominently argued that embedded wh-questions are implemented in phrase structure by a covert complementizer she terms *WH*. She shows that *WH* is subcategorized for by verbs, as shown in the following example.

(32) a. * We believed whether he was there.

b. We inquired whether he was there.

(33) a. We believed that he was there.

b. * We inquired that he was there.

(Bresnan, 1970, 303-304)

That *WH* is a complementizer is shown by its complementary distribution with other complementizers.

(34) * I know that whether he came.

(35) * For whom to own a rifle doesn't affect me.

(36) * It doesn't matter to them whether that you march.

(37) * I asked for what John to do.

(Bresnan, 1970, 311)

Later debate over classification of clause embedding verbs uncovered predominantly semantic phenomena which further subdivided clause-embedding verbs into

several natural classes: P-selecting verbs, Q-selecting verbs, and verbs such as 'know' which select for either P-clause or Q-clause continuations.[3] Hintikka (1976) showed that a class of verbs termed 'factive' verbs, such as 'know' in 39, can take either P or Q-type complements.

(38) a. Albert claimed that the supervisor dismissed the employee.

b. * Albert claimed whether the supervisor dismissed the employee.

c. * Albert claimed if the supervisor dismissed the employee.

d. * Albert claimed who the supervisor dismissed.

e. * Albert claimed who dismissed the employee.

f. * Albert claimed when the supervisor dismissed the employee.

(39) a. Albert knew that the supervisor dismissed the employee.

b. Albert knew whether the supervisor dismissed the employee.

c. Albert knew if the supervisor dismissed the employee.

d. Albert knew who the supervisor dismissed.

e. Albert knew who dismissed the employee.

f. Albert knew when the supervisor dismissed the employee.

(40) a. * Albert wondered that the supervisor dismissed the employee.

b. Albert wondered whether the supervisor dismissed the employee.

c. Albert wondered if the supervisor dismissed the employee.

d. Albert wondered who the supervisor dismissed.

[3]The terminology P-selecting and Q-selecting originate from the syntax literature are correspond respectively to -Q and +Q. Herein, we use the terms P/Q-selecting in reviewing the syntactic literature and -Q/+Q when discussing the experiment.

e. Albert wondered who dismissed the employee.

f. Albert wondered when the supervisor dismissed the employee.

The ability of *know* to select either P or Q is unlikely due to lexical ambiguity, as P and Q clauses can be coordinated and embedded by *know*.

(41) John knows/revealed/guessed who left early and that Mary was disappointed. (Groenendijk and Stokhof, 1983, 66)

Hintikka (1976) posits that a verb selects for both P or Q continuations if and only if it is factive; that is, the verb entails its Propositional complement, as seen in 43.

(42) a. John knows that Bill ate the chocolate $\models$ Bill ate the chocolate

b. John knows whether Bill ate the chocolate.

(43) a. John claims that Bill ate the chocolate $\not\models$ Bill ate the chocolate

b. * John claims if Bill ate the chocolate.

Henceforth, debate over the Factivity Hypothesis centered around whether *know*-verbs or their complements were syntactically or semantically distinct from pure P or pure Q-selecting predicates. Munsat (1986) proposed that *know* P-complements and Q-complements are headed by a distinct *wh-that* complementizer.

(44) a. I wonder where John went (wh-Q)

b. I know where John went (wh-that)

(45) a. I know that John went home (wh-that)

b. I believe that John went home (that)

(Munsat, 1986, 191)

Lahiri (1991) and Berman (1991) differ in whether *wonder* vs. *know* complements are type-theoretically distinct. Lahiri (1991) proposes that the embedded wh-elements in 44a and 44b are both question-typed [4]. Berman (1991) proposes that while *wonder* class verbs truly select questions, verbs such as *know* select only propositions, and that such verbs involve a covert operator which type-shifts the embedded question. This is corroborated by morphological evidence from Basque, which uses a specific complementizer -(e)na for factives (Adger and Quer, 2001).

(46) Ikusi dot [asko-rik ez dakia-na]
seen AUX.1SGE.3SGA much.PART not know.3SGE.3SGA-COMP
I have seen/realized that he doesn know much.

(47) Ezagun da [kopiatu daua-na]
known be.3SGA cheated AUX.3SGE.3SGA-COMP
It is clear/known that he cheated (on the exam).

Ginzburg (1995a) and Ginzburg (1995b) argue convincingly that the respective interrogative complements of *wonder* vs *know* are in fact truth-conditionally and

[4] defined following Karttunen (1977) as sets of propositions ($< t, t >$)

type distinct. On the basis of Quantificational Variability evidence, he proposes that intensional question-embedding verbs such as *wonder* take a true question as their complement, but for extensional question-embedding verbs such as *know*, this complement is coerced by an operator (I) into a fact, which in his ontology is distinct from a proposition. Likewise, he proposes that the that-complements of *know*-type verbs are not in fact propositions, but propositions coerced into facts with another operator (D). His analysis though unifies the explanation of the Factivity Hypothesis and QVE.

Munsat (1986) observed that *wonder*, but not *know*-class verbs license Negative Polarity Items, as seen in 49.

(48) a. * I know how he ever did it.

b. I wonder how he every did it.

c. I don't know how he ever did it.

(49) a. * I know why anybody bothers to listen to him.

b. I wonder why anybody bothers to listen to him.

c. I don't know why anybody bothers to listen to him. (Munsat, 1986, 67)

Moreover, Adger and Quer (2001) demonstrate the ability of some typically P-selecting predicates to embed what they term Unselected Embedded Questions in traditional NPI-licensing contexts of negation or questionhood.

(50) a. * Julie admitted/heard/said if the bartender was happy?

(51) a. Did Julie admit/hear/say if the bartender was happy?

b. Julie didn't admit/hear/say if the bartender was happy.

(Adger and Quer, 2001, 110)

5.1.2 wh-Islandhood

Grammatical islands such as 52a and 52b have been amongst the core data for transformational approaches to grammar since Ross (1967). Insofar as such structures are surprisingly unacceptable for native speakers, they have been taken to yield insights into constraints on the transformational component. Because they are sensitive to linguistics-internal considerations such as the P/Q feature on CP-taking verbs, island constraints were argued by Ross (1967),Chomsky (1986), and Pesetsky (1987), to reflect native-speaker grammatical competence. Some strong islands are in fact quite exceptionless; Ross (1967)'s Coordination constraint among them. Empirically, it is quite possible that both performance and competence phenomena are conflated in the term islandhood. One of the difficulties for any exploration of islandhood is the possibility of a family of phenomena having multiple causes. Sociologically, it may be difficult for any nuanced picture to emerge that integrates ideas from the syntactic, semantic, and psycholinguistics literature; moreover, it is methodologically difficult to conduct confident investigation in the psycholinguistics of islands if syntax and semantics of islands are not pinned down. That said, the phenomena themselves may prove helpful to settling debate: strong islands are so inviolate that they suggest a syntactic explanation, whereas weak

islands are ameliorable and gradient, suggesting semantic and/or psycholinguistic explanations.

(52) a. Who did Albert say that they dismissed?

b. # Who did Albert wonder whether they dismissed?

c. ? Which employee did Albert wonder whether they dismissed?

Weak islandhood presents a critical battleground over the grammatical status of islandhood. Advocates of the purely structural approach to islandhood put forth several arguments for their view. First, weak islandhood varies cross-linguistically; on the widespread view of Miller and Chomsky (1963) that the parsing module is cognitively universal, and on the predominant Principles and Parameters linguistic theory of the time, such structures dovetailed readily with the purely grammatical approach. Second, the selectivity of weak islands is engendered by linguistic entities which possess no obvious processing cost on a Derivational Theory of Complexity (Fodor and Garrett, 1975). Advocates of the performance view, on the other hand, point out the gradience of weak islands and the difficulty of writing down a comprehensive theory of island selectivity; the literature since Ross (1967) has pointed to a myriad number of exceptions and ameliorations which defy a single unifying consensus approach to these phenomena. For instance, Pesetsky (1982)'s approach derives the acceptability of movement via an elsewhere condition; extraction is acceptable if the extractee is either properly governed (in its base position) or antecedent governed (at intermediate and surface positions). Proper government requires roughly that the extractee be subcategorized for (so that accusative case objects will always be properly governed, and adjuncts never). Antecedent

government holds when the extraction path meets certain well-formedness requirements; for any extraction path, it must not cross any other extraction path (all paths must be properly contained). Notably, this would ban cross-serial dependencies and restrict the grammar to context free power; it would also eliminate movement-movement ambiguity.

5.2 Proposal

We model how our +Q ambiguity hypothesis derives the Alexopoulou and Keller (2007) experimental finding that +Q contexts in English produce degraded Magnitude Estimates (Cowart, 1997) when compared to -Q controls. This experimental finding supports acceptability judgment-based findings in the syntax literature, namely, the island status of interrogative complementizers, found in Ross (1967).

We test whether a Minimalist Grammar which encodes the greater diversity of +Q continuations would result in greater processing effort of island sentences by computing the entropy of parse forests (Hale, 2006) in +Q continuations metric. Hale's (2006) Entropy Reduction Hypothesis suggests that the reduction of parse forest entropy triggered by a new word measures the cognitive load that an incremental parser exerts to disconfirm predictions no longer congruent with the unfolding sentence. The ERH models the total amount of work required to parse a sentence like 'Albert wondered who the supervisor dismissed' as the summed decreases in entropies of parse forests situated on prefixes 'Albert *', 'Albert wondered *'..., where

* represents the suffix language of grammatically licensed continuations.

We find that the +Q condition exhibits greater mean entropy reductions **(25.3 b)** than the -Q condition **(12.5 b)**, deriving acceptability judgments reported in the literature on *whether*-islands (Bresnan, 1970; Chomsky and Lasnik, 1977; Pesetsky, 1987). The ambiguity hypothesis of islandhood eliminates representational overhead by deriving the processing difficulty of islands as a direct consequence of the +Q/-Q hypothesis. Additionally, our account readily derives the fact that 'D-linking' or specification of the extracted DP ameliorates islandhood (Pesetsky, 1987; Hofmeister and Sag, 2010). When the extraposed DP is specified (*which employee* in 28c as opposed to *who* in 28b), the parser can utilize s-selection clues such as animacy to guide the parse towards more likely predictions, resulting in reduced ambiguity.

5.3 Test

The Ambiguity Hypothesis suggests that Islands exhibit greater incremental ambiguity than Non-islands due to the +Q Ambiguity operant in Islands. To formalize the respective roles of competence and performance in this prediction, we adopt two linking hypotheses. On the competence side, we implement Bresnan (1970) using the Derivational Minimalist Grammar formalism (Stabler, 1997). On the performance side, we adopt the Entropy Reduction (Hale, 2006) complexity metric. The ERH operationalizes the linguist's notion of ambiguity by modeling the intrinsic sentence processing work required to parse a word as the amount of in-

formation uncertainty reduced by the word. The comprehender's analytical state is modeled as a weighted intersection of the competence grammar and the unfolding sentence. This weighted intersection is itself a probabilistic grammar whose weights correspond to degrees of belief attached to competing parses of the unfolding sentence; information-processing work has occurred when weight accrues toward certain analyses.

A Minimalist Grammar inspired by Bresnan (1970)

We implement a Derivational Minimalist Grammar (Stabler, 1997) which distinguishes between +Q and -Q complementizers, deriving the paradigm in Example 27. This grammar constitutes an analysis of the problem space encountered by the human parser on the sentence in Alexopoulou and Keller (2007), so the amount of work required to incrementally parse this problem space can be calculated.

=C S	**claim** V-Q	=>V v	**who** D -wh
=T C	**wonder** =P V+Q	**that** =T CC	**about** =D P
=Q C	**wonder** =CQ	**whether** =T CQ	**with** =D P
will =v-Q =D T	V+Q	**if** =T CQ	**why** P -wh
=v-Q =D T	**wonder** =Q V+Q	=T +wh Q	**how** P -wh
will =v+Q =D T	**wonder** =CC	**do** =T +wh Q	**when** P -wh
=v+Q =D T	V+Q	**must** =T +wh Q	**where** P -wh
V-Q =CC V-Q	=>V+Q v+Q	**fire** =D V	T << P
V-Q V-Q	=>V-Q v-Q	**fire** V	
claim =CC V-Q	=>V-Q v-Q	**Mary** D	

Figure 5.1: Sample Unweighted Minimalist Grammar

As seen in 5.1, the grammar implements Bresnan (1970)'s central observation that there exist two natural classes of embedded complement phrases (-Q embedded

ABOUT = [lemma = wonder & pos = VVD|Z][lemma = about]
WHETHER = [lemma = wonder & pos = VVD|Z][lemma = whether]
IF = [lemma = wonder & pos = VVD|Z][lemma = if]
Q = [lemma = wonder & pos = VVD|Z][(lemma = wh.* & (pos=W.* | pos =IN | pos = RB)) | lemma = how]
THAT = [lemma = wonder & pos = VVD|Z][lemma = that & pos = IN]

Figure 5.2: Corpus Queries for complements of 'wonder','claim' on NYT

declaratives, +Q embedded interrogatives) and that clause-embedding verbs can pattern differently with respect to which complements they select. For example, verbs such as *wonder* select only -Q continuations, whereas verbs such as *think* select +Q continuations. The grammar also implements filler-gap wh-question structures and verb transitivity ambiguity.

5.3.1 Weighted Minimalist Grammars

The Entropy Reduction Hypothesis requires a probabilistic formal grammar which intersects a theory in the form of a grammar with the problem space of an experiment. We implement this problem space via a weighted corpus which we parse up into a Minimalist mini-treebank.

Utilizing the weighted mini corpus methodology from Hale (2006), we built a weighted training corpus which reflects statistical subcategorization facts for the verbs in the experiment. Using the CQP (Christ, 1993) queries shown in Fig. 5.2, we collected counts from *New York Times* corpus (Sandhaus, 2008) for different complements subcategorized for by *wonder* and *claim*. We also obtained counts for embedded verb token transitivity and adjunct taking. The training corpus

utilizes a factorial design, such that each sentence varies across four parameters: Matrix Verb (*claim* vs. *wonder*); Complement Type (*if*, *whether*, embedded question, *that*, *about*); Embedded Verb Transitivity (transitive, intransitive); and Adjunct/Argument extraction. All but the last parameter was estimated as described above; the Adjunct/Argument parameter was set at 0.5. The minicorpus weights each sentence by the products of normalized counts for each parameter, so that the weighted corpus represents an intersection of relevant +Q/-Q and *wh*-island parsing facts.

+Q wonder	**sentence frame**	**-Q claim**
0.00720	"who does Jane wonder/claim about"	0.00047
0.01240	"who does Jane wonder/claim if Mary will punish"	3.1965e-5
0.00150	"who does Jane wonder/claim if Mary will punish with"	3.8877e-6
0.01240	"who does Jane wonder/claim if Mary will punish Mary with"	3.1966e-5
0.01334	"who does Jane wonder/claim who will punish"	8.47745e-5
0.01334	"who does Jane wonder/claim who will punish with"	8.47745e-5
0.00314	"who does Jane wonder/claim when Jane will punish"	1.99785e-5
0.00038	"who does Jane wonder/claim when Jane will punish with"	2.4298e-6
0.00314	"who does Jane wonder/claim when Jane will punish Jane with"	1.99785e-5
0.00314	"who does Jane wonder/claim where Jane will punish"	1.99785e-5
0.00038	"who does Jane wonder/claim where Jane will punish with"	2.4298e-6
0.00314	"who does Jane wonder/claim where Jane will punish Jane with"	1.99785e-5
0.00314	"who does Jane wonder/claim how Jane will punish"	1.99785e-5
0.00038	"who does Jane wonder/claim how Jane will punish with"	2.4298e-6
0.00629	"who does Jane wonder/claim how Jane will punish Jane with"	1.99785e-5
0.00314	"who does Jane wonder/claim why Jane will punish"	1.99785e-5
0.00038	'who does Jane wonder/claim why Jane will punish with"	2.4298e-6
0.00314	""who does Jane wonder/claim why Jane will punish Jane with"	1.99785e-5
0.00019	"who does Jane wonder/claim that Mary will punish"	0.01710
2.3326e-5	"who does Jane wonder/claim that Mary will punish with"	0.00208
0.00019	"who does Jane wonder/claim that Mary will punish Mary with"	0.01710
0.00677	"who does Jane wonder/claim whether Mary will punish"	0.0
0.00081	"who does Jane wonder/claim whether Mary will punish with"	0.0
0.00671	"who does Jane wonder/claim whether Mary will punish Mary with"	0.0

Figure 5.3: Sample Minicorpus for Embedded Verb *punish*

We parsed the weighted mini corpus using mcfgcky, obtaining a weighted Minimalist Grammar treebank, and computed entropies for the prefix parse forests for the test sentences.

5.4 Results

We used *mcfgcky* (Grove, 2010) to train on the above minicorpus and test on the Alexopoulou and Keller (2007) sentences. We tested whether a Minimalist Grammar which encoded the greater diversity of +Q continuations would result in greater processing effort of wh-islands using the Entropy Reduction metric (Hale, 2006). We obtained greater total entropy reductions for the Island condition than for the Control condition, as seen in Fig.5.4.

	Entropy Reduction Summed Across the Sentence
+Q	$\sum H \downarrow = 3.379$ bits
-Q	$\sum H \downarrow = 1.899$ bits
ΔQ	1.480 bits

Figure 5.4: Total Entropy Reductions with Argument/Adjunct Parameter at 0.50/0.50

Almost all of the difference is due to the +Q ambiguity. As seen in Fig. 5.5, Island and Control conditions exhibit almost the same Filler-Gap ambiguity, but the Island condition exhibits the additional +Q ambiguity. This is verified through

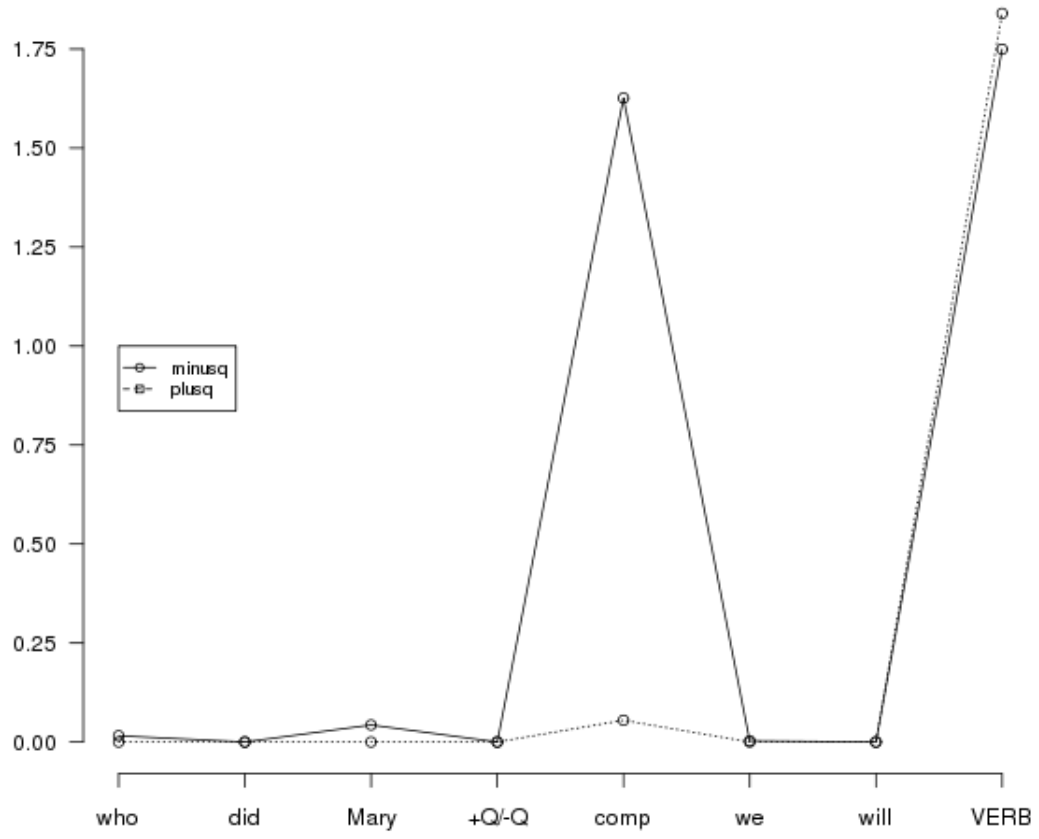

Figure 5.5: Incremental Entropy Reductions for Island and Control Conditions

a post-hoc analysis of incremental parser state output immediately following the matrix verb, so that the probabilistic parse forest corresponding to "Who did Mary wonder/claim *" was obtained. From this parse forest, we extracted the MG category which corresponds to a matrix verb phrase with a nominal gap ($v_{wonder} - q$ for Island, $v_{wonder} - q$ for Control). This category was found to have different entropy and branch weightings depending upon the experimental condition. Island condition matrix verb phrases ($v_{wonder} - q$) have 2.556 bits of total entropy, whereas

Control condition matrix verb phrases ($v_{claim} - q$) have 0.011 bits of total entropy. However, this difference in entropy is not just due to the encoding of complementizers in the grammar, nor the simple statistical distribution of sentences, but is rather due to multiple dimensions of the +Q ambiguity. The Island condition spreads probability across three different branches, as seen in Fig.5.6, and the one-step rewriting entropy of these branches is 1.290 bits. The particular branch which corresponds to embedded *wh*-questions is encoded with Q -q. This category itself has 2.328 bits of entropy, contributing 1.222 bits of entropy to its parent. Our interpretation of this fact is that the main result is driven by such hierarchical uncertainties stemming from syntactic complexity.

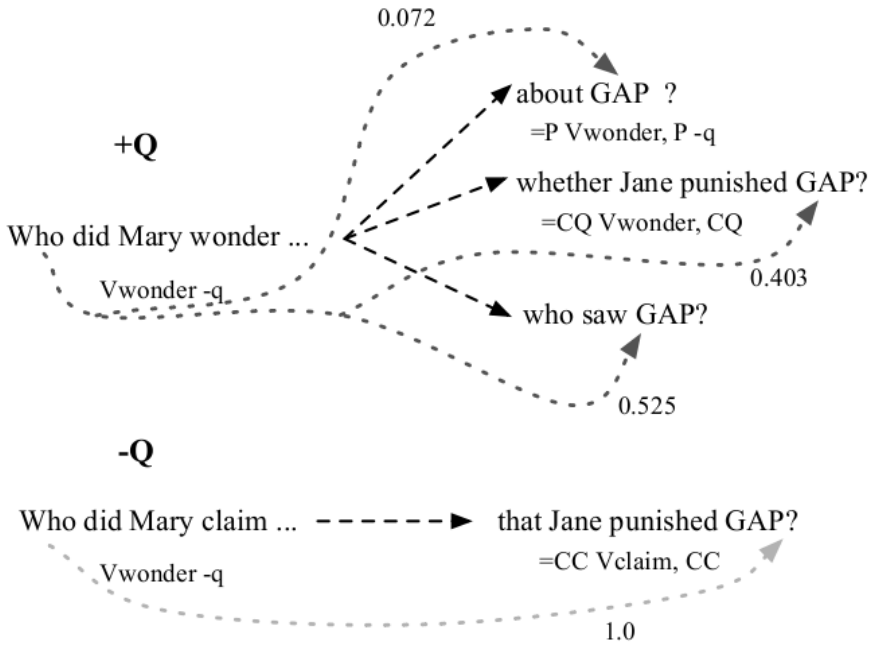

Figure 5.6: Branch Probabilities for Island and Control Conditions After Verb

In formulating the experiment, implementing the minicorpus methodology for wh-

islands posed a particular set of challenges which steered us towards conservative modeling choices, potentially dampening the effect. Recall that in the minicorpus methodology, the experimenter implements a factorial parameter table and constructs a sentence for each entry in the table, then weights each sentence by the product of the parameterization of that sentence. The strong independence assumptions in the factorial design and the resulting minicorpus means potentially modeling borderline grammatical or ungrammatical sentences as training data, potentially with artificially high probabilities. Take for example the below weighted sentence from the minicorpus:

0.00314 "who does Jane wonder/claim how Jane will punish"

The above probability for the sentence is likely artificially high, but this probability comes directly from multiplying the independent parameters from the factorial design. This probability could be more brought in line with reference corpora by modeling the co-occurrence between +Q, embedded verb transitivity, and complementizer with an interaction term, but the resulting term itself would need to be estimated from co-occurrences in reference corpus data. In our study, wh-extraction and *wonder*-class verbs were generally rare events, so depending on modeling co-occurrences of +Q and particular lexicalized facts was not a workable methodology.

Conversely, in the course of experimental design we made several conservative decisions that stemmed directly from the minicorpus methodology applied to wh-islands. In contrast to garden path sentences, which are uncontroversially a per-

formance phenomenon, islands as a whole represent a set of phenomena that can be highly difficult to pin down as structural, semantic, or psycholinguistic. The Miller and Chomsky (1963) framework and the minicorpus methodology assume that 1) the set of processable sentences is a proper subset of the grammatical set of sentences, and that 2) the parser is guided by the grammar to only compute grammatical analyses of sentence. Why this represents a difficulty for modeling weak islands is that the experimenter needs to assemble a set of training sentences to weight by parameters in reference corpora. That means that the experimenter is required to model the set of relevant grammatical continuations twice: intensionally, in the grammar; and extensionally, in the corpus. For many island phenomena sentences, this poses the experimenter with a dilemma: declare a marginal sentence to be grammatical, but unprocessable, and write it down in the minicorpus with a non-zero probability, or declare a marginal sentence to be ungrammatical, and not in the minicorpus. In modeling the entropy of weak islands, we have to commit to including some weak-island sentences in the minicorpus to define the analysis space, but neither theory nor corpus methodology give the experimenter support to cherry-pick which sentences are ungrammatical, and should not be trained on, versus which sentences are grammatical, and are to be included with the weight indicated by the parametrization. For example, an early version of the experiment included a grammar and a corpus which attested extraction from subject, but after consideration, the Subject-Island constraint was taken to be a fact of the grammar, and therefore subject-extraction was removed from the analysis.

5.5 Discussion

We report in the following section some background results which we argue are congruent with the account developed in this work.

5.5.1 *wh*-Islandhood is Combinatorial in Nature

Kluender and Kutas (1993) found that embedded interrogatives impose more cognitive load than embedded declaratives even when no extraction from the embedded sentential takes place. They crossed a Question-type factor with two conditions, a Yes/No Question control Condition and a *wh*-Question condition by an embedding factor with three conditions: a *claim that* condition, a *wonders whether* condition, and an embedded question (*wonder who*) condition. Subjects participated in a neurolinguistic EEG task and a speeded acceptability judgment task. The Yes/No Question Condition controls for extraction out of the embedded sentential while preserving the interrogative mood. As seen in Fig.5.7, subjects in the neurolinguistic EEG task exhibited greater deflection in ERP in the *wonders whether* and the *wonders who* conditions than the *claim that* condition, even when extraction out of the embedded sentential did not occur. Kluender and Kutas (1993) propose that these results indicate separate, discrete correlates for *wh*-processing and island-structure building. This effect straightforwardly follows from the +Q Ambiguity (Bresnan, 1970), which our model capitalizes on.

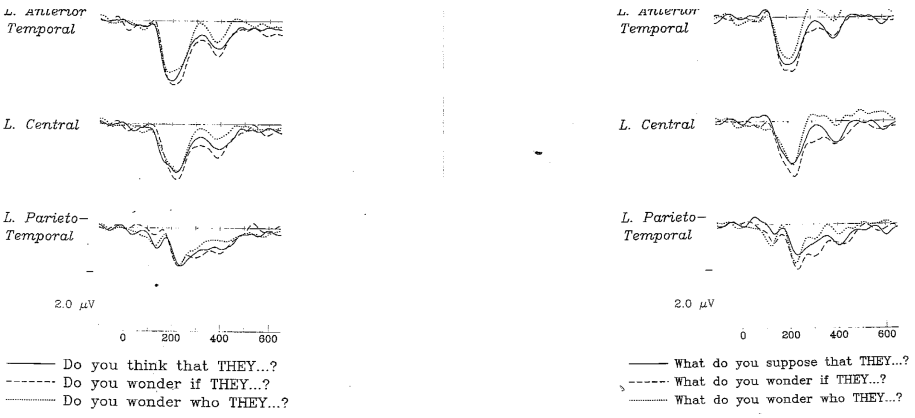

Figure 5.7: (Kluender and Kutas, 1993)

5.5.2 *wh*-Islandhood is Ameliorated by Extragrammatical Factors

Hofmeister and Sag (2010) found that potential islands with 'Complex' extraposed NP arguments (such as in 53a) are processed more quickly than Simple cases (such as in 53b).

(53) a. Which employee did Albert learn whether they dismissed after the annual performance review? (p.30)

b. Who did Albert learn that they dismissed after the annual performance review? (p.30)

That this clearly extragrammatical level of treatment remedies the *wh*-island effect suggests that the proper explanation of the effect itself is extragrammatical and therefore a performance artifact. Hofmeister and Sag (2010) suggest that the effect may be due to working memory, but do not fully outline how.

We argue that the *wh*-island effect is due to the intrinsic work required of the parser by the structure-assigning task at hand. The asymmetry in the amount of intrinsic processing work required by the grammar originates from a structural ambiguity which is present in the Island condition but not the Control condition. Because our account predicts that the *wh*-island effect is the sum of the +Q ambiguity and the Filler-Gap ambiguity, it predicts that savings in the ambiguity budget for either ambiguity should yield easier processing. Hofmeister and Sag (2010)'s effect would then fall out as a case where +Q ambiguity is operant but Filler-Gap processing is ameliorated.

5.5.3 *wh*-Islandhood is not a Working Memory Phenomenon

However, Sprouse et al. (2012) found no significant correlation between working memory capacity and the ability to process *wh*-islands. Neither serial recall nor n-back capacity predicted perceived acceptability on a variety of island processing tasks, including *wh*-islands. Given the tendency of acceptability judgments and on-line processing measures to correlate, this result would belie a working memory basis for the *wh*-island effect. Thus Sprouse et al. (2012) suggest, contra Hofmeister and Sag (2010), that island processing is a grammatical, not a processing phenomenon. The results of Hofmeister and Sag (2010) and Sprouse et al. (2012) would appear at first glance to be irreconcilable.

As our account offers ambiguity rather than working memory as an explanation for the *wh*-island effect, it is not falsified by the results of Sprouse et al. (2012).

We point out that the hierarchical processing of nested structures such as *wh*-islands bears little resemblance to the n-back and serial order working memory tasks employed in Sprouse et al. (2012), which are predominantly tests for short term capacity for memorization of discrete items.

5.5.4 Further Discussion

An additive-factors approach (Sternberg, 1969) to empirical experiments could tease apart the complexity of *wh*-island processing. The modeling methodology of this paper integrates with such an empirical program in a way which provides cohesion to the myriad number of factors involved in islandhood. It also provides for a possibly more informative set of results than appealing to the modularity of the grammar or the processor as the sole province of islandhood. One possible avenue of exploration would be to expand the repertoire of +Q and -Q verbs commonly used in experiments. Island experiments are almost entirely limited to the most common matrix verbs, such as *wonder* (+Q), 'claim' (-Q), and *say* (ambiguous); Hofmeister and Sag (2010) is a notable exception in that it uses twelve verbs with ambiguous +Q/-Q status, but it lacks a control condition with pure -Q verbs. By compiling experimental results where conditions utilize verbs across the +Q/-Q continuum, verb ambiguity provides us with a new independent variable we can modulate in experiment. Further work would also endeavor to dispense with the minicorpus methodology, which as noted present special problems with respect to the problem of modeling complex phenomena of borderline grammaticality.

Chapter 6

Reduced Relative Garden-Pathing and the Unaccusativity Hypothesis

This chapter provides a new account of the asymmetrical effect of argument structure on garden pathing difficulty. (54a) with unaccusative *melted* is easier to comprehend than (54b) with unergative *raced*.

(54) a. **Unaccusative:** EASY The butter melted in the oven was lumpy.

b. **Unergative:** HARD The horse raced past the barn fell. (Stevenson and Merlo, 1997)

Verbal ambiguity contributes to the difficulty of such reduced relative clause (RRC) garden paths. When the embedded verb is obligatorily transitive (55a), RRC disambiguation is less difficult than in optionally transitive conditions (55b), because it renders the misleading active voice, intransitive analysis of the ambiguous substring untenable (Pritchett, 1992; MacDonald, 1994). [1]

[1]We use the term 'reanalysis' in this paper to simply mean the effort required to successfully integrate the disambiguating token into the parse, and abstract away from whether the parsing architecture is serial, parallel, or somewhere in between.

(55) a. The ruthless dictator captured in the coup was hated throughout the country.

b. The ruthless dictator fought in the coup was hated throughout the country. (MacDonald, 1994)

Stevenson and Merlo (1997) appeal to the lexicon for an explanation of this processing asymmetry, positing that causative v applies lexically for unaccusatives, but syntactically for unergatives. We argue instead that unaccusatives and unergatives are both made causative "in the syntax". We explain the extra difficulty of unergative causation by appealing to the causative-PP co-occurrence restriction noted in Hoekstra (1988),Levin and Rappaport-Havov (1995), and Folli and Harley (2006): unergative causation requires an argument-attached prepositional phrase (PP).

(56) a. **The window broke.**

b. Pat **broke the window.**

(57) a. **The soldiers marched** *(to their tents.)*

b. The general **marched the soldiers** **(to the tents)*[2]. (Levin and Rappaport-Havov, 1995)

Wayne Harbert (p.c.) points out that temporal adverbials can also license unergative causation for English native speakers.

[2]Bold and italics added for emphasis.

(58) a. **The soldiers marched** *(all day.)*

b. The general **marched the soldiers** **(all day)*[3].

Hoekstra (1988) and Folli and Harley (2006) show that adjunct PPs do not license unergative causation. Spatial PPs occurring with unergative manner of motion verbs (such as *float*, below) are potentially ambiguous between an directional (argument) and a locative (adjunct) reading.

(59) The boat floated under the bridge (Zubizaretta and Oh, 2007, 28).

This ambiguity is indeed syntactic; Hoekstra (1988) and Zubizaretta and Oh (2007) show that Dutch auxiliary selection is sensitive to this ambiguity. With intransitive unergatives in Dutch in perfect aspect, the *zijn* ('be') auxiliary forces the directional reading of the PP, the *hebben* ('have') auxiliary the atelic locative reading. Unergatives with argument attachment of PP pattern together with unaccusatives in being easily causativized.

These data seems to suggest two key observations: that telicity is required for productive causative alternation, and that the prepositional phrases attached to unergative manner-of-motion verbs are potentially ambiguous between atelic locative Prepositional Phrases and telic Path Phrases. Thus, for manner-of-motion unaccusative verbs such as *break*, causative alternation is totally productive, whereas for manner-of-motion unergatives such as *race*, causative alternation depends upon analyzing the prepositional phrase as a Path Phrase and not a Locative PP.

[3]Bold and italics added for emphasis.

However, in parsing unergative RRCs, the verbal ambiguity and the PP-attachment ambiguity both precede the disambiguating token. The co-occurrence restriction between argument-PPs and the causative forms of unergative verbs means that the parser encounters greater uncertainty in the unergative case, which we hypothesize explains the greater difficulty of unergative RRC:

Hypothesis: PP-attachment ambiguity contributes to the greater difficulty of unergative RRC via a co-occurrence restriction between unergative causativization and argument-attachment of PP.

On their Radical Construction Grammar-based account, McKoon and Ratcliff (2003, 2005) posit that unergative RRC are ungrammatical rather than unparseable. Contra McKoon and Ratcliff (2003, 2005), the account of the reduced relative asymmetry herein honors the Miller and Chomsky (1963) methodology in taking the form of an explicit parsing architecture whose components are a memory component, a control structure (parsing module), and the grammatical knowledge that the parser uses to inform decisions. We assume no special role of memory in this account, and abstract away from it henceforth. As a proxy for the control structure, we employ Total Information to model the effort spent to disambiguate reduced-relative clause garden path sentences.

We argue that the grammatical knowledge used by the parser in the reduced relative task to be the knowledge of acceptable causative forms for unergatives and unaccusatives argued for in Levin and Rappaport-Havov (1995), Hoekstra (1988), and Folli and Harley (2006). We formalize this grammatical co-occurrence restriction

using Stabler (1997)'s mildly context sensitive Minimalist Grammars (MG) formalism. In our grammar, we assume a constructional view of the lexicon, adopting the Distributed Morphology (DM) framework for verbal argument structure.

As we showed in Chapter 3, entropy reduction does not predict garden pathing, as it predicts garden paths and main clause continuations to be equally hard. We therefore predict that surprisal, but not entropy reduction, will show both unergative and unaccusative reduced relatives to be more difficult than their respective main clause continuations. We predict that Total Information will show this as well.

Moreover, we predict that Total Information will be greater in the unergative reduced relative than the unaccusative reduced relative. We first predict that Entropy Reduction will be greater for the unergative reduced relative, as the reduced relative ambiguity will inherit the prepositional phrase ambiguity as associated by the causation-PP co-occurence. We also predict the Surprisal will be greater for the unergative reduced relative, because the unergative reduced relative clause requires two uncommon events (reduced relative clause, Path-attachment of the prepositional phrase), while the unaccusative reduced relative clause requires only one uncommon event (reduced relative clause) as both Path-PP and Location-PP attachments allow for a causative analysis of the VP, and therefore a reduced relative.

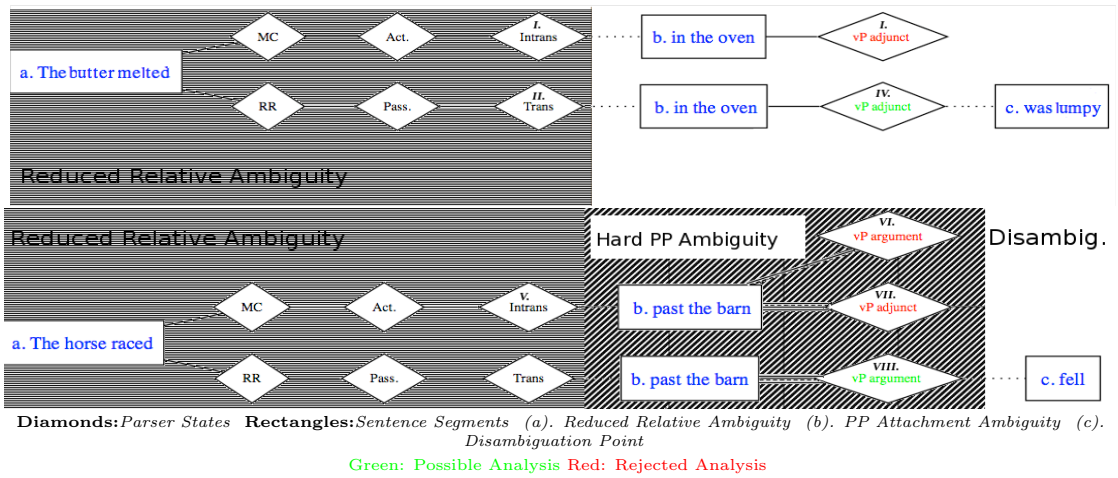

Figure 6.1: Relationships Between Reduced Relative and PP-Ambiguity

6.1 Grammar

To compute surprisals and entropies of a lexical-grammatical event, we require a tractable grammatical formalism that allows for the definition of a lexical grammar and can be probabilistically weighted. We constructed the below MG analysis of filler-gap structures and lexical structure of the pertinent verbs.

Roots & PP-Attachment	**Passive**	**Reduced Relative Clause**
::=P =D SC	::=T C	the::⇒ agrD D
::=SC $=\surd_{do}$ V_{delta}	::=voice T	::=Crel +nom agrD
::=D $=\surd_{delta}$ V_{delta}	::=V_{do} T	::=T +arel Crel
::=$\surd_{do}$ =D V_{do}	the :: =N D -k;	::=pass +prel Crel
melted::$\surd_{delta}$	the :: =N D;	::=Opass T
walked::$\surd_{do}$	::=V_{delta} +k T	who::=N D -k -arel
fell::$\surd_{do}$	::=V_{delta} =D voice	which::=N D -k -arel
past::=D P	was::=V_{delta} +k Opass	::=N D -k -prel
in::=D P	were::=V_{delta} +k Opass	who::=N D -arel
D << P	::=V_{delta} +k pass	which::=N D -arel
$\surd_{do}$ << P	horse::N -nom	::=N D -prel
$\surd_{delta}$ << P	butter::N -nom	

Figure 6.2: MG of reduced relative clauses and argument structure

The features which govern Move similarly consist of Licensor and Licensee features. An item with a licensor feature selects an item with the corresponding Licensee feature, to create a derived item whose phonetic feature is some concatenation of the childrens', and whose head is the category of the Licensor.

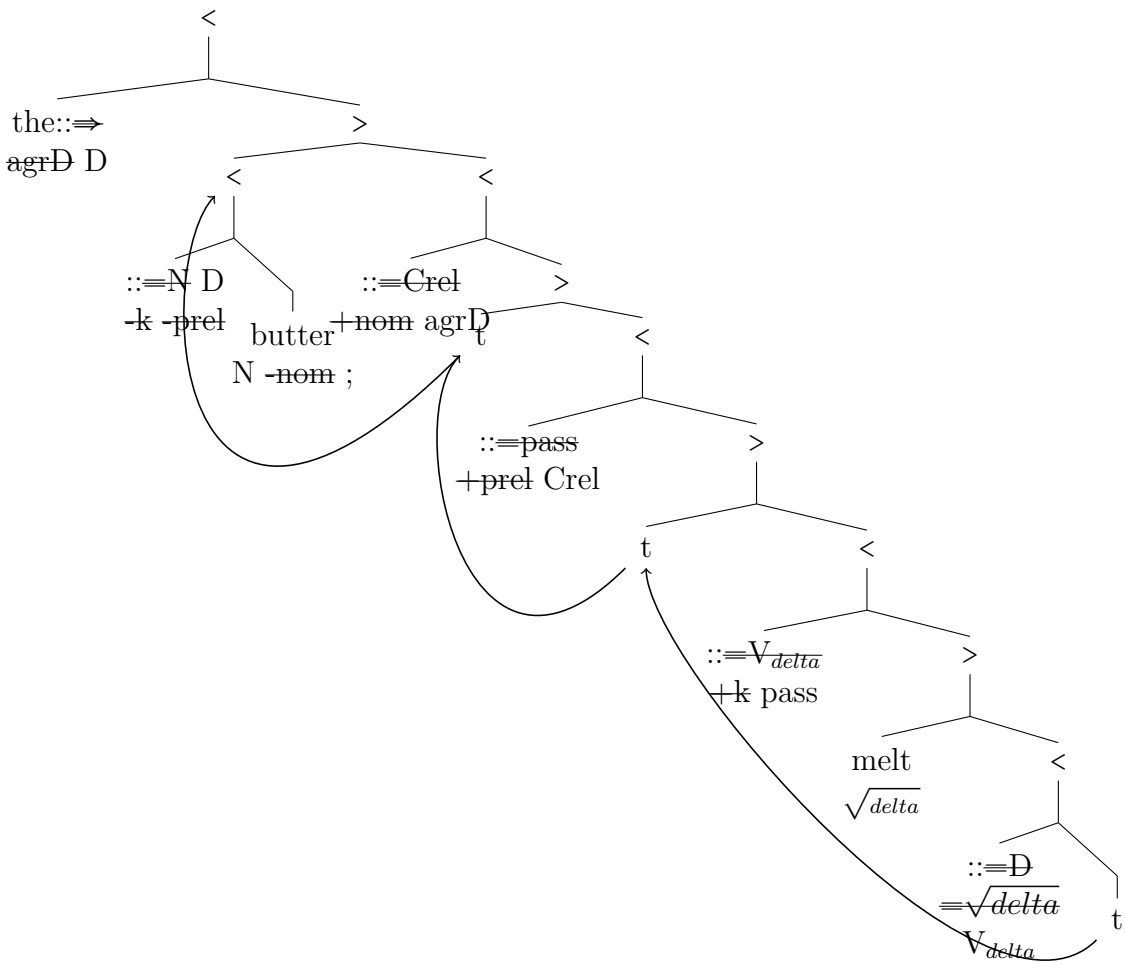

Figure 6.3: Derived Fragment for Unaccusative RRC 'The butter melted in the oven'

While Minimalist Grammars are not context free grammars, they share a useful property with the mildly context sensitive formalisms they are equivalent to: they possess a context-free backbone which can be weighted and estimated as a PCFG

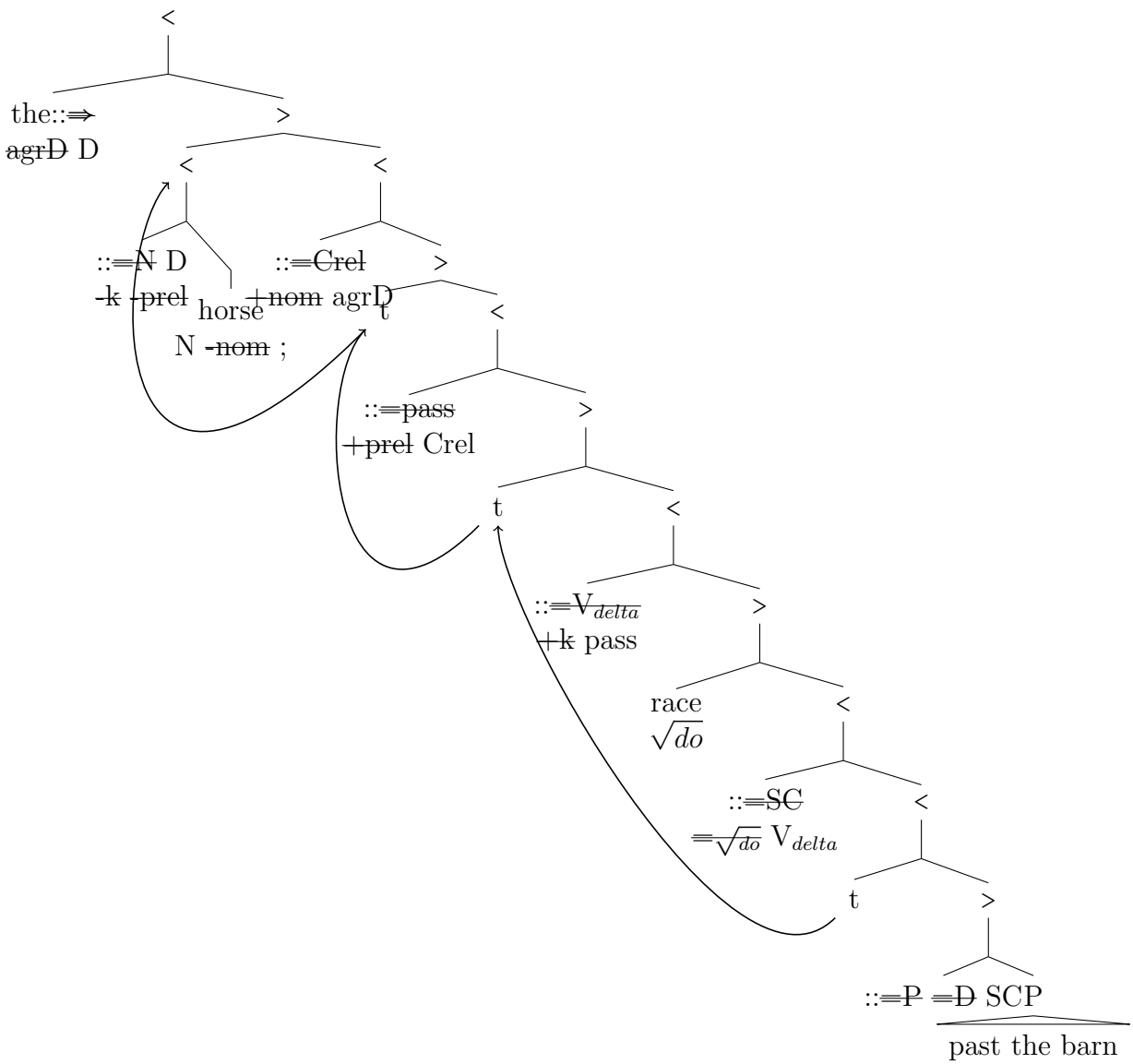

Figure 6.4: Derived Tree for Unergative RRC 'The horse raced past the barn'

can. Parsers which compute the conditional entropy of mildly context sensitive formalism do exactly this, and treat the rules of Minimalist Grammars and other mildly context sensitive formalisms as context-free rules with more complex string yield functions than simple concatenation.

6.2 Methodology

We developed a Stablerian Minimalist Grammar to operationalize our hypothesis that co-occurrence restrictions on causation were responsible for the more severe garden pathing of unergatives reported in Stevenson and Merlo (1997). This Minimalist Grammar formalizes the co-occurrence restriction on the causative alternation of unergative verbs reported in Hoekstra (1988),Levin and Rappaport-Havov (1995), Folli and Harley (2006) and Zubizaretta and Oh (2007). The correspondence between directed motion unergatives and unaccusatives was rendered with a Small Clause category (Hoekstra, 1988; Folli and Harley, 2006) which selects a Distributed Morphology-style Root, while standard motion unergatives are treated as a simple root structure. The causative v-head selects the Small Clause, so that passives and causatives of unaccusatives and directed-motion unergatives are possible, but causatives and passives of unergatives with no small clause are not possible. We also formalized the promotion analysis (Kayne, 1994) of relative clauses, with the reduced-relative relative clause co-occurrence with passivization. RRC are generated in our grammar via a covert relative pronoun which selects only a covert passive morpheme. This captures the distribution in English of the RRC construction.

We employed an MG parsing system which uses the Guillaumin compiler as a front end into an intermediate (MCFG) formalism. This system generates a probabilistic model by using a training and a testing phase. At training time, the parser parses a mini-corpus of full sentences, and maintains counts of how many times a particular rule was used. The parser uses the Weighted Relative Frequency estimation of

Chi (1999) to estimate a backbone PCFG from these counts. The sentences in the training mini-corpus are representative of sentence forms which are possible disambiguated forms for the conditions the experimenter is interested in. For the task at hand, the training corpus presented each verb form in both a reduced relative frame and a main clause frame. Only verbs were lexicalized; all other categories are not.

In this methodology, the testing mini-corpus contains the sentences whose conditional entropies the experimenter is interested in. The PCFG which was estimated at training is renormalized to fit the testing sample (rules which are used in training but not in testing are factored out, and other probabilities are adjusted similarly), and the parser computes from the training sample the right-hand side vector of local entropies and the fertility matrix ala Grenander (1967) for each sentence prefix.

A sample of 24 reduced relative garden path sentences (12 unaccusative; 12 unergative) from the Stevenson and Merlo (1997) study was obtained. From this sample, training and testing corpora were constructed. Sentences were controlled for length. The locally ambiguous segments of sentences, which consisted in each case of the sample as a reduced relative sentence fragment consisting of the sequence D NP V P NP, was only edited by simplifying, equiprobably balancing, and length-controlling NPs. The verb tokens of interest were not altered. The disambiguating segments were edited for length and to control for passivization, with 6 active voice and 6 passive voice disambiguations for each condition. We weighted the active voice disambiguations relative to reduced relative disambiguations using a param-

eter estimated from version 3 (BNC XML Edition), reflecting the concept that the parser is uncertain between reduced relative disambiguations and main clause disambiguations but has knowledge regarding the much greater likelihood of main clauses. From this manipulated sample, a 48 sentence training corpus was created, where each sentence was represented twice: once with just the locally ambiguous segment as a main clause, and once with the complete garden path sentence. The 24 sentence sample of complete garden path sentences in the manipulated sample served as the testing mini-corpus.

6.3 Results

Table 2	*Unerg. ER*	*Unerg. S*	*Unacc. ER*	*Unacc. S*
the	0.0	0.0	0.0	0.0
children/butter	0.0	0.0	0.0	0.0
walked/melted	1.372	0.795	1.217	1.476
through/in	0.180	0.056	0.921	0.302
the field/saucepan	0.0	0.0	0.0	0.0
disappeared	0.872	5.801	0.104	4.833
.	0.535	3.036	0.535	3.036

Figure 6.5: ER and Surprisal Results by Condition

The unergative condition elicits significantly greater reduction in entropy and surprisal from the locally ambiguous segment to the disambiguated terminal, correctly deriving the result that unergative reduced relative processing is more difficult for human subjects. The analysis also derives the prediction that garden-path continuations will have greater difficulty than main clause continuations, as the surprisal of the reduced relative clause event is greater than the main clause continuation. The greater total information at the disambiguating token in the unergative case

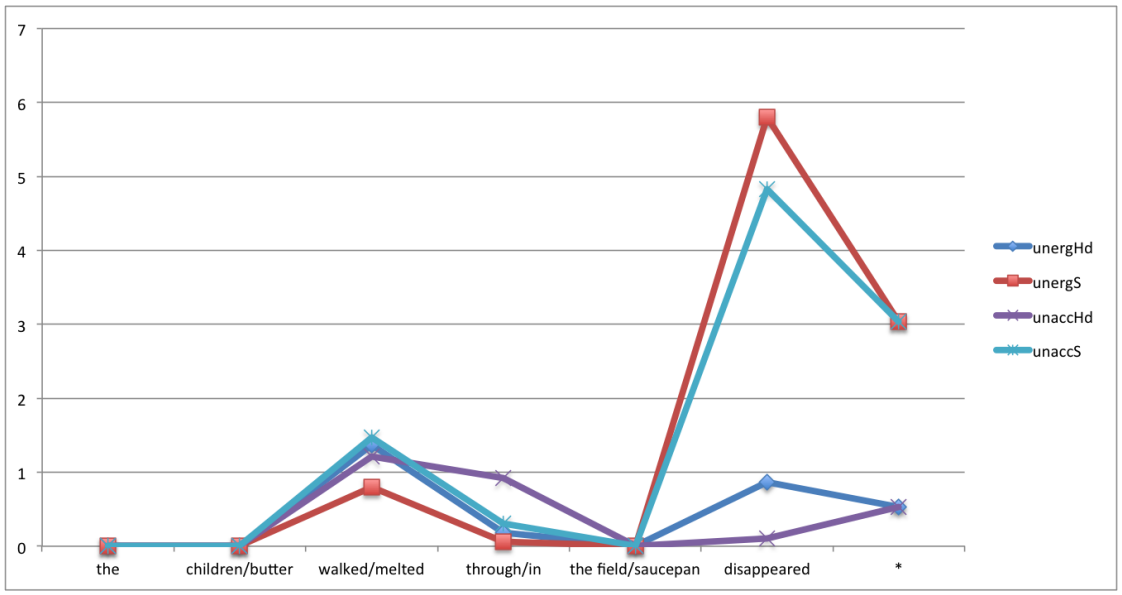

Figure 6.6: Entropy Reduction and Surprisal for Unergative and Unaccusative Reduced Relatives

stem from the fact that the RRC event and the prepositional phrase attachment event are coupled in the unergative case, but not in the unaccusative case. In the unaccusative case, the parser does not need to eliminate a prepositional phrase attachment in order to proceed, whereas in the unergative case, the parser must eliminate the adjunct attachment of PP, as the globally correct RRC parse depends upon an argument attachment of the PP.

Surprisals and entropies from the model are generally low, because the underlying grammar is mostly unlexicalized and there are few degrees of freedom. That there can be surprisal or entropy reduction at a particular token on our analysis follows from these facts, because the syntactic event of N following a determiner has probability one on our simple grammar that excludes adjectives, and the lexical identity of the noun does not enter into the analysis.

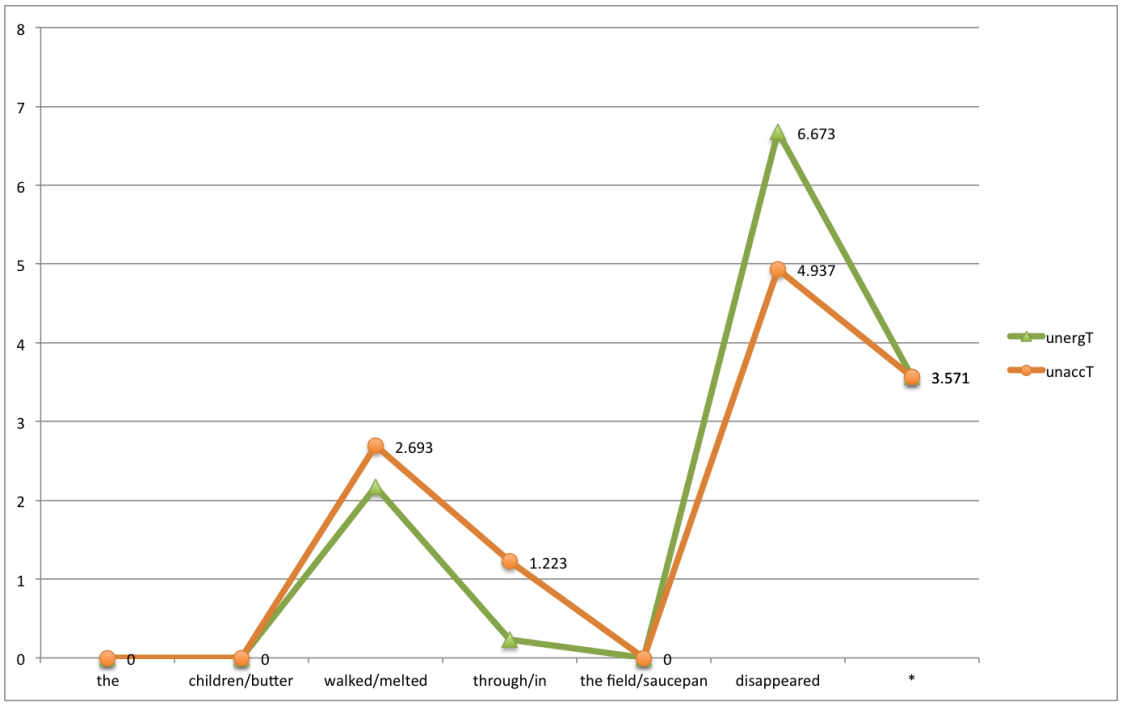

Figure 6.7: Total Information for Unergative and Unaccusative Reduced Relatives

6.3.1 Discussion

Our approach attempts to be maximally parsimonious while according with several widely-held methodological assumptions regarding the modularity of the human sentence processor and the human language faculty. First, following Miller and Chomsky (1963), we employ a processing model where the set of parsable sentences is a subset of the set of grammatical sentences. This commits us to the view that, barring independent evidence to the contrary, both the unaccusative and unergative reduced relatives are grammatical. Second, we respect the Competence Hypothesis (Miller and Chomsky, 1963), which argues that the grammar used by the human sentence processor is the same grammar that linguists study. Our performance hypothesis is simply our MG grammar and the co-occurrence restriction it encodes. Third, we adopt the Distributed Morphology (Halle and Marantz,

1993) framework for encoding argument structure. Distributed Morphology takes the view that the Lexicon is not modularly encapsulated, but is itself derived. Finally, we take the view that ambiguity resolution is the main end towards which the human sentence processor works. To this end, we adopt as a parsing model the Entropy Reduction Hypothesis, which models parsing difficulty as the reduction in analytic entropy from one sentence token to the next.

Our account argues that reduced relatives are grammatical but difficult to process. Uncontroversially, reduced relative garden path sentences require that a noun-verb-prepositional phrase sequence which initially appears as an active voice, intransitive main clauses must ultimately be analyzed as passivized, transitive, reduced relative constructions. The co-occurrence restriction observed in the literature is that unaccusatives freely participate in the causative alternation, but motion unergatives require a directional, argument-attached PP. We have proposed that the processing asymmetry for unergative and unaccusative reduced relatives is situated precisely on the production asymmetry for unergative and unaccusative verbs presented in Hoekstra (1988),Levin and Rappaport-Havov (1995), and Folli and Harley (2006). In the unaccusative reduced relative condition, the reduced relative ambiguity can be resolved independently of the prepositional phrase attachment ambiguity, as unaccusatives can be reanalyzed as causative independent of PP attachment. However, in the unergative case, the reduced relative ambiguity and the PP attachment ambiguity must be resolved together, since causative reanalysis of unergatives requires a PP attachment. Transitivity reanalysis of unergatives requires argument attachment of the prepositional phrase in order to realize the verb as a passive, causative, reduced relative.

The study confirms that the co-occurrence restriction between prepositional phrase attachment and causation of unergatives is responsible for the increased processing difficulty of unergative reduced relatives. The unergative condition exhibits a later, more severe reduction in entropy than the unaccusative condition. Notably, these results obtained even though we abstracted away from the relative frequency of the reduced relative and main clause constructions. A replication of this experiment where these constructions are weighted realistically would undoubtedly exacerbate the processing asymmetry predicted in our model. It would also be desirable in future work to estimate the parameter for adjunct versus argument attachment of prepositional phrases in general from corpus; some techniques for this estimation are discussed in Merlo and Ferrer (2006).

We have attempted to analyze the contribution of verb type to reduced relative processing with maximum parsimony. Though the subject matter is inherently concerned with the interaction of syntactic factors with lexical factors, we wanted to explore this interaction while stipulating as little as possible about the nature of the lexicon. Thus, we used the Distributed Morphology framework to model a well reported co-occurrence restriction (Hoekstra, 1988; Levin and Rappaport-Havov, 1995; Folli and Harley, 2006) in the literature. Our Minimalist Grammar implements this co-occurrence restriction, but does not render a stance on what is done "in" or "out" of the lexicon. Empirically, this account of the effect in Stevenson and Merlo (1997) dovetails neatly with Pritchett (1992); while both unaccusatives and unergatives are optionally transitive, the transitivity of unergatives is conditional on prepositional phrase attachment. The unergative condition is biased against transitive reanalysis; we would predict unaccusatives to be as dif-

ficult as Pritchett's optionally transitive verbs, but more difficult than obligatorily transitive verbs.

Though garden path phenomena have been studied since Bever (1970), it is only with the recent convergence of work in lexical semantics and syntax that psycholinguists have sought to classify garden path effects by the lexical semantics of the reanalyzed verb. MacDonald (1994) found an alleviating effect of obligatorily transitive, but not optionally transitive, verbs in the processing of reduced-relative garden path constructions. O'Bryan (2003) interprets this result as evidence that lexical information can trigger a internal argument expectancy in the parser, but is mostly agnostic on how this expectancy will guide parsing. In the same work, O'Bryan examined the effect of verb telicity on reduced relative clauses in garden path construction, and offers the following:

Event Structure Processing (ESP) hypothesis: During comprehension, event structure information, accessed in a verb's lexical entry, affect parsing decisions. If the verb is inherently telic, the verb will be parsed as having an underlying direct object (p. 29)

The account avoids special appeal to lexical modularity, and assumes a strongly incremental sentence processing mechanism common to both unergative and unaccusative conditions. The account proceeds with the strong assumption that all analyses are maintained as long as they are congruent with the incoming sentence, and the long-held view that sentence processing is primarily an ambiguity resolution task. Our account readily accommodates not only the Stevenson and Merlo

(1997) data, but more generally the Pritchett (1992) results of a processing advantage for obligatorily transitive verbs over optionally transitive verbs in the reduced relative construction. The obligatorily transitive verbs are never congruent with an intransitive main clause analysis, whereas optionally transitive verbs are congruent with both the intransitive main clause analysis and the transitive, passivized, reduced relative analysis, until the disambiguating token is met.

Chapter 7

Conclusion

This thesis pursued two main questions: 1) how can lexical structure help explain sentence processing effects; and 2) what do surprisal and entropy actually measure. In part, the first question is a device to help explicate the second; subcategorization tends to produce relatively 'crisp' grammatical judgments and is easy to model empirically from corpora, so it serves as a device towards examining the difficult question of what the linking theories of surprisal and entropy actually represent.

The argument laid out in this thesis is that a statistically weighted lexicon represents an optimization performed by the human sentence processor to solve a worst-case exponential-time problem with average time linear performance. The probabilistic lexicon enables a weighted-beam search that employs advantages of both serial and parallel processors. The weighted-beam search can allocate disproportionate processing resources to a favored analysis while maintaining resources for other possible analyses. In this respect, the probabilistic lexicon constitutes an optimal betting strategy for gambling resources for the gain of sentence processing time, empirically optimized for the average case. The parser works in close parallel with the syntax and the lexicon to insure linear time processing; where lexical optimization preconditions are violated by laboratory sentences, they result in a parsing experience that is supra-linear and cognitively painful. The thesis ar-

gues that surprisal and entropy measure distinct types of processing breakdowns resulting from the failure of this optimization. Because the human sentence processor has limited resources, the parser can encounter processing inefficiency when encountering highly entropic and highly surprising parser states.

We argue that there are two general types of parsing breakdown: either the parser is overly slavish in committing processing resources to the globally incorrect parse (surprisal effects, Type I), or it is overly hesitant to commit processing resources to any one parse (entropy effects, Type 2). We argue that most sentence processing sits between these two extremes and encounters small amounts of both surprisal and entropy in the day-to-day processing of ambiguity. In the Type 1 surprisal case, the parser has grown overconfident in a parse which turns out to be globally incorrect; subjectively, the parser is *surprised* at some event in the sentence processing stream such that it cannot recover easily. The raw cost of a surprisal effect maintains on a generally parallel view of the parser having too do much reranking work to incorporate a very surprising token (Levy, 2008), or on a limited parallel view of the parser having lost the globally correct parse out of its beam (Boston et al., 2008). In either case, sentence processing difficulty arises because the parser has dedicated too many parsing resources to a disconfirmed parse. We argue that this type of breakdown covers such phenomena as garden path sentences, filled gap effects, and other plausibility effects more generally.

On the other hand, in the Type 2 Entropy case, the parse is overly *hesitant* to commit processing resources to any one parse. The parser does not encounter enough clues early in the sentence to guide the parse efficiently to one conclusion. Rather,

each encountered token yields new ambiguities. The parser reaches a ceiling at which point it cannot maintain the full breadth of parses under analysis. In a highly entropic parser state, the sentence processor allocates processing resources equiprobably to many different hypotheses; in essence, it is not optimizing for any one case. Such cases arise, we argue, in many of the classic cases of sentence processing breakdown for alleged working memory reasons, such as center-embedded sentences and wh-island effects; in such cases, we argue that it is the parser's hesitation over multiple nesting hypotheses that results in parsing breakdown. We argue that this mode of processing breakdown explains most sentences classified as difficult for working memory reasons, and that sentences difficult for working memory reasons are best modeled by entropy. The thesis argues that cases such as weak island effects and center embedded sentences are difficult because the parser runs out of working memory to allocate to the maintenance of one complex analysis. We argue that with respect to such measures as acceptability judgment, the predictions of entropy and entropy reduction on such sentences are generally equivalent, as the sentence's distinguishing phenomena is the peak entropy at a moment of maximal sentence difficulty, and the sentences generally end with entropy zero, in the unambiguous case.

This overall view is congruent with recent views of working memory as a limit on how many sub-problems can be independently solved, as opposed to a limit on how many items can be retained in a stack. In general, the thesis takes the stance that the parser regularly resolves minor ambiguities with no complaint. From this viewpoint, the thesis argues that that multiple cases of sentences processing difficulty (wh-islands, unergative reduced relative clause garden path sentences)

are better viewed as combinations of ambiguities rather than singularly difficult cases. It follows from this viewpoint that many ameliorations are not the special cases they are made out to be in the sentence processing literature. In the case of islands, -Q-selecting verbs do not directly ameliorate islandhood, as there is nothing special about +Q-selecting clauses in isolation. Rather, for wh-islands, it is the build up of ambiguities (+Q/-Q, filler-gap, verb transitivity) that cumulates to a point at which the parser cannot cope. Similarly, unaccusative reduced relative clauses do not give rise to the garden path effect, not because unaccusativity makes garden patching *artificially easier*, but because the attested dependence of unergative verbs on prepositional phrase attachment makes unergative RRC *artificially difficult* to parse. Again, the parse confronts reduced relative, transitivity, +Q/-Q ambiguities on a regular basis, but it is only the compounding of two or more ambiguities in a particular way that gives rise to sentence processing breakdown.

Moreover, for ameliorations that might legitimately constitute special cases, we argue that the mechanism for amelioration is by cues which indicate movement preference, reducing global ambiguity. In particular, we argue throughout that this analysis of sentences thought to be problematic for working memory parsimoniously explains many ameliorative empirical effects that must be stipulated on other accounts, as these ameliorative cases involve phenomena reducing the ambiguity of the sentence. For weak islands, the phenomenon that 'd-linking' ameliorates whether-islands is readily explainable on our account as an animacy-subcategorization effect; the identity of the 'd-linked' object as an animate or inanimate DP reduces movement entropy on a probabilistic theory of subcatego-

rization that has verb token-animate DP correspondences as a parameter. For center embedded sentences, the Gibson (2000) effect that pronouns ameliorate center-embedded sentences falls out from the Ambiguity Hypothesis when we consider that subject hood of pronouns is an identifying cue reducing movement entropy.

Throughout, we argue that this view of the parser and the lexicon promotes maximum parsimony and reproducibility of scientific efforts. Above all else, this thesis attempted to eradicate special appeal to the lexicon as an explanation for sentence processing asymmetries. Rather, the thesis has adopted the methodology of incorporating independently attested lexical-structural co-ocurrences as explanations for psycholinguistic phenomena in a general framework of sentence processing. We seek no special mechanisms in either the lexicon or the parser. In this framework, the parser is a general processing architecture which takes specifications from the syntax, empirically weights the parse space from experience, and allocates processing resources to candidate parse in proportion to its best determination at each point in the sentence. Neither the lexicon nor the sentence processor have any special escape hatches to accommodate special appeal; instead, our theories take the form of lexical entries weighted empirically from corpora, and the linking theories are theory-agnostic complexity metrics from information theory. The ultimate hope for this project is for a deep explanation of how the parser utilizes the lexicon, and to argue that the proper place of lexical semantics in sentence processing is not as a general escape hatch for sentence processing theories, but as a source of optimizations by which the parser makes possible its day-to-day job of crunching ambiguity.

Chapter 8

Appendix: mcfgcky

The experiment described in the **wh-Islandhood** Chapter of the project utilizes a Minimalist Grammar (Stabler, 1997) parsing system to compute entropies and surprisals. This appendix describes the implementation of a statistical Multiple Context Free Grammar (MCFG) (H. Seki and Kasami, 1991; Nakanishi et al., 1997) parser implemented in OCaML for use as a backend to such a system. This system utilizes the mg2mcfg compiler described in Guillaumin as a front end. The parser features wild-card parsing over unknown segments of arbitrary unknown length, after Lang (1988), for use over probabilistic grammars as part of a psycholinguistic modelling tool for computing the entropies (Hale, 2006) and surprisals (Hale, 2001) of expressive grammars intersected with automata.

8.0.2 Multiple Context Free Grammars

Multiple Context Free Grammar (H. Seki and Kasami, 1991) is a mildly-context sensitive (Avarind K. Joshi and Weir, 1992) formalism, as are Minimalist Grammar (Stabler, 1997), Combinatory Categorical Grammar (Steedman, 2000), and Tree Adjoining Grammar (Joshi, 1985).

Multiple context free grammar rules (and more generally, mildly context sensi-

tive grammar rules) differ from context-free rewriting rules in delineating abstract syntax from concrete syntax. Abstract syntax refers to our method of rewriting nonterminals as other non-terminals, whereas concrete syntax refers to how we manage the string yields of our symbols.

In a context-free production,

$$\alpha \rightarrow \beta\gamma \tag{8.1}$$

α, β and γ represent strings, and the rewriting operation denoted by $\rightarrow$ conflates a category-forming operation and a concatenation operation. The category-forming operation and concatenation operation are disassociated in mildly context sensitive grammars since the basic unit is non-concatenate (think of a derived tree in TAG, which permits adjunction into it), and concrete syntax permits fancier string yield operations than simple concatenation.

Multiple Context Free Grammar rules operate over tuples of strings. MCFG rules are of the form

$$A_0 \rightarrow f[A_1, A_2, ...A_q] \tag{8.2}$$

where the function f takes as arguments tuples of strings and returns an A_0 which

is also a tuple of strings.

MCFG rules are often notated in practice with the following notation of Albro, which makes clear the evaluation of f.

$$t7 \rightarrow t4\ t5\ t6\ [(2,0)][(1,0);(0,0);(1,1)] \tag{8.3}$$

This example can be read off as follows. An index (x, y) refers to the yth member (in 0-initial list notation) of the x argument. The semicolons represent concatenations, whereas the brackets designate members of the yield tuple. Thus, in the above example, a new tuple of category t7 is formed, where the first member is simply the 0th member of the $t6$ tuple, and the second member is formed from the concatenation of: the 0th member of $t5$; the 0th member of $t4$; and the 1st member of $t5$. A rule such as the above example, where the second member of $t7$ is formed from the interpolation and subsequent concatenation of members of $t4$ and $t5$, can thus be seen as similar to a movement operation (MG), wrap rule (CCG), or adjunction instance (TAG).

8.0.3 MCFG String Parsing

We present a sample MCFG derivation. In 8.1, we present a small MCFG which captures noun phrase movement for unaccusative verbs such as 'fell'.

S	→	VP	$[(0,1);(0,0)]$
VP	→	V NP	$[(0,0)][(1,0)]$
NP	→	D N	$[(0,0);(1,0)]$
V	→	'fell'	
N	→	'boy'	
D	→	'the'	

Figure 8.1: Sample MCFG

9	S	$[Frag_{(0,3)}]$	→ 8
8	VP	$[Frag_{(0,2)}, Frag(2,3)]$	→ 6,7
7	NP	$[Frag_{(0,2)}]$	→ 4,5
6	V	$[Frag_{(2,3)}]$	→ 3
5	N	$[Frag_{(1,2)}]$	→ 2
4	D	$[Frag_{(0,1)}]$	→ 1
3	'fell'	$[Frag_{(2,3)}]$	→ Term
2	'boy'	$[Frag_{(1,2)}]$	→ Term
1	'the'	$[Frag_{(0,1)}]$	→ Term

Figure 8.2: Sample MCFG Derivation

The VP rule in this MCFG implements the desired mapping between 'deep structure' and 'surface structure' by seperating 'abstract' and 'concrete' syntax. The VP rule's abstract syntax creates a VP symbol by taking the NP as the 'complement' of V, but the string yield function over VP evaluates as the tuple ("fell", "the boy"). The S rule is therefore able to 'move' this NP up in its concrete syntax, as seen in the sample derivation in 8.2.

In our sample derivation we included backpointers in the Parsing as Deduction style (Shieber et al., 1995); we can retrieve a derivation by following backpointers into subderivations. Importantly, the derivation is just a grammar (Billot and Lang, 1989) conditioned on the full string; this fact will enable us to port analytical tools from traditional PCFGs over incremental parse states.

8.0.4 Prefix Parsing as Intersection of (M)CFG and Finite State Automaton

Billot and Lang (1989), and Lang (1988) introduced the shared packed parse forest for representing a potentially infinite number of derivations of a CFG prefix parse. The shared packed parse forest follows from the Hillel et al. (1960) proof that context free grammars are closed under intersection with finite state automata; The shared packed forest represents an infinite number of prefix parses as the intersection of a context-free grammar with a finite state machine. The resulting shared packed forest is itself a recursive context free grammar, obtained by sharing parent nodes as children of multiple rules, and by sharing multiple contexts in single contexts.

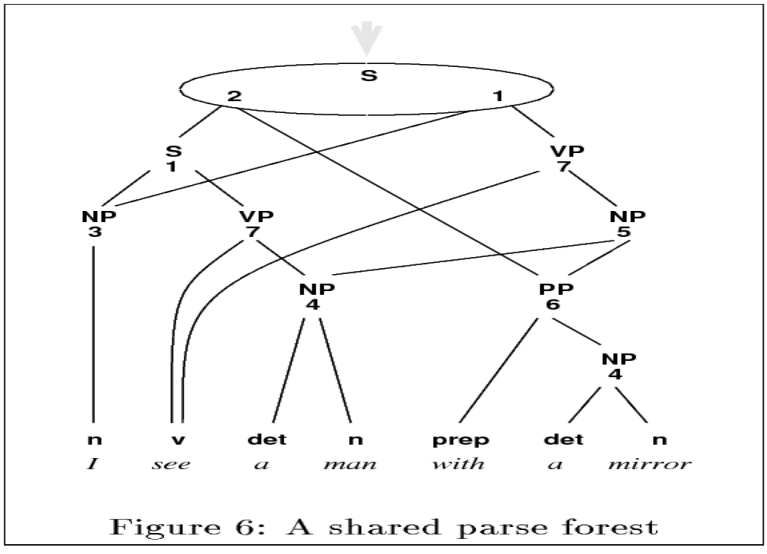

Figure 8.3: Billot and Lang (1989); Lang (1988): Ambiguous Parse, as Shared Parse Graph

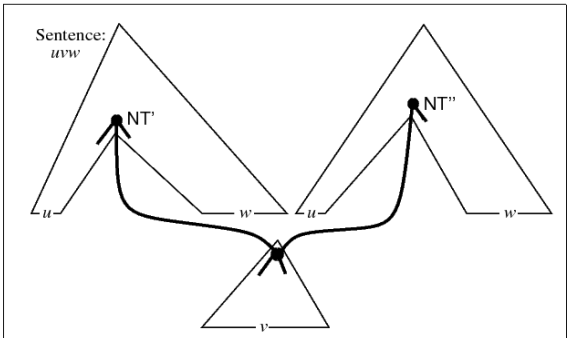

Figure 7: Two parses sharing a subtree

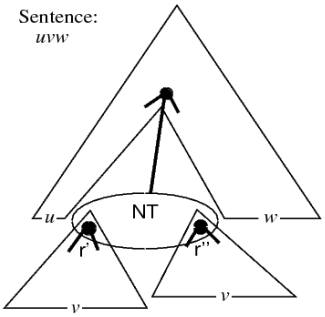

Figure 8: Two parses sharing a context

Figure 8.4: Billot and Lang (1989); Lang (1988): Ambiguous Parse, as Item/Tree Sharing

9	S	$[Frag_{(0,2)}]$	→ 8
8	VP	$[Frag_{(0,2)}, Wild(2)]$	→ 6,7
7	NP	$[Frag_{(0,2)}]$	→ 4,5
6	V	$[Wild_{(2)}]$	→ 3
5	N	$[Frag_{(1,2)}]$	→ 2
4	D	$[Frag_{(0,1)}]$	→ 1
3	'*'	$[Wild_2]$	→ Term
2	'boy'	$[Frag_{(1,2)}]$	→ Term
1	'the'	$[Frag_{(0,1)}]$	→ Term

8.0.5 Probabilistic Parsing and MCFG

PCFGs have the following properties (Manning and Schütze, 1999)

- Place Invariance: The probability of a subtree does not depend on where in the string the words it dominates are.

- Context Free: The probability of a subtree does not depend on words not dominated by the subtree.

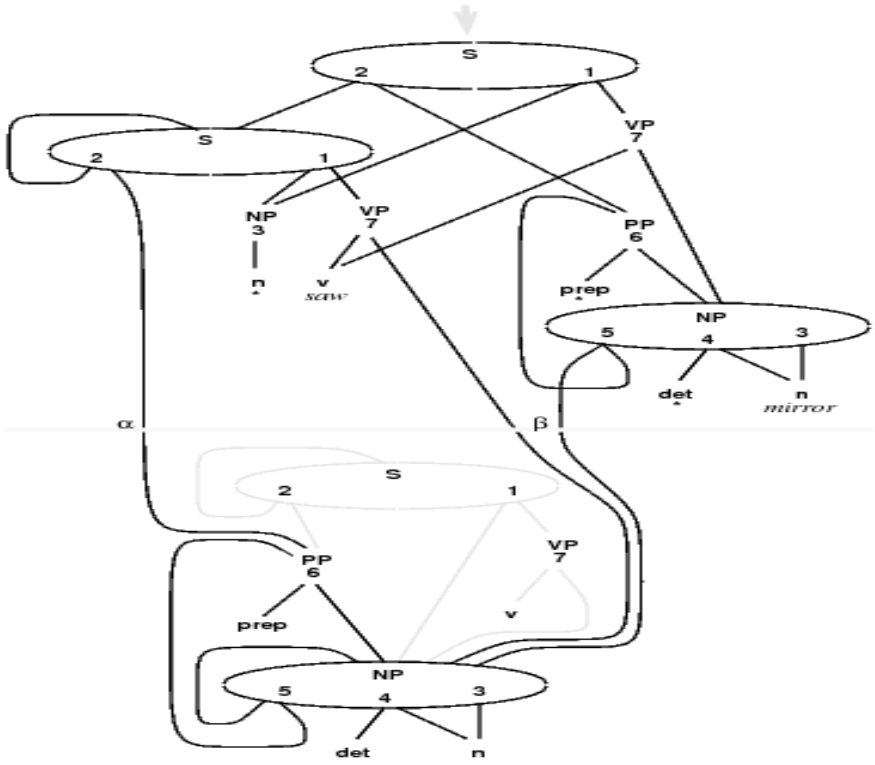

Figure 9: Full parse forest for an incomplete sentence

Figure 8.5: Intersections of Context Free Grammars and Automata

- Ancestor Free: The probability of a subtree does not depend on nodes in the derivation outside the subtree.

Our MCFG derivation trees have a context free 'abstract syntax' (H. Seki and Kasami, 1991; Kallmeyer, 2010). The MCFG abstract syntax in fact enjoys Place Invariance, Context Freeness, and Ancestor Freeness via its "context-free backbone" (Avarind K. Joshi and Weir, 1992), permitting the extension of PCFG methods to more expressive mildly context sensitive grammars. The parameters for productions in a probabilistic mildly context sensitive grammar can thus be estimated via (Weighted) Relative Frequency Estimation for PCFG (Chi, 1999),

as shown below.

$$P(A \rightarrow \xi) = \frac{\sum_{i=1}^{n} f(A \rightarrow \xi; \tau_I)}{\sum_{i=1}^{n} f(A; \tau_I)} \tag{8.4}$$

Relative Frequency Estimation estimates the likelihood of an outcome by taking a count of the number of outcomes in a data set, and dividing by a count for all contexts in which the event could have had that outcome. For a PCFG G, the event of rewriting a parent A will have outcomes which are rules in G with left hand side A. To estimate the probability of a rule $A->\xi$ from corpus τ, we only need count the number of occurrences of $A \rightarrow \xi$ in τ as the numerator, and divide by the total number of instances of A as left hand side of a rule in τ.

A PCFG with arbitrary probabilities on productions may fail to define a probabilistic language. For a PCFG G to define a probabilistic language, we require that P be *proper* and *consistent*[1]. Following Chi (1999), a PCFG is *proper* iff $\sum_{\lambda:A\rightarrow\xi)\epsilon G} P(A \rightarrow \lambda) = 1$, i.e. for any nonterminal A, the probabilities on rules rewriting A sum to 1. A PCFG G is *consistent* if $\sum_{x\epsilon\Sigma^*} P(S \Rightarrow x) = 1$, if the set of all strings derived by G have probabilities summing to 1.

That G is proper is not sufficient to ensure that G defines a probabilistic language. Stolcke (1995) demonstrates a PCFG which is proper but fails to define a probabilistic language, as shown in 8.6.

[1]Following here the terminology of Stolcke (1995),Chi (1999), Hale (2006), and not the terminology of Jelinek and Lafferty (1991), who uses *consistent* to term what we mean by *proper*.

2/3 S → S S
1/3 S → 'a'

Figure 8.6: Inconsistent PCFG

Stolcke (1995) shows that the set of all sentences defined by this probabilistic grammar have probabilities which sum to greater than 1. Put loosely, the above PCFG generates nonterminals 'faster' than those nonterminals can be cashed out into terminals. A given PCFG is said to be *proper* if for all nonterminals X, the rule probabilities with parent X sum to 1. This is insufficient to insure that a probabilistic grammar is *consistent*, that is, that it defines a probabilistic language. A given PCFG G is consistent iff all string probabilities add to 1, which is the case if the PCFG's fertility matrix M is invertible and if the fertility matrix's [2] spectral radius (largest eigenvalue) ≤ 1.0. The spectral radius of M shows how recursive the PCFG is; if $i(G) \leq 1.0$, then a derivation will halt with certainty. If $i(M) \geq 1.0$, then non-terminals are being introduced 'faster' than they are rewriting into terminals, with the disastrous result that some strings gain infinite probability.

Chi (1999) shows that if a PCFG is trained from a treebank with (Weighted) Relative Frequency Estimation, it is guaranteed to be consistent. The fertility matrix A is indexed by nonterminals, in which the (i, j) entry in A records the expected number of the nonterminal j from one rewriting of the nonterminal i. Grenander (1967) and Stolcke (1995) show that the inversion $(I - A)^{-1}$ gives the transitive closure of the fertility relation; this is crucial for the computation of PCFG entropy and requires a consistent PCFG.

[2] equivalently, momentum matrix, first-moment matrix

8.0.6 Probabilistic Intersection of a (M)CFG and Finite State Automaton

Inside Probability

We will need (for several reasons) to compute the inside probability of a category situated on a string or finite state automaton. Following Manning and Schütze (1999, 392), we define the inside probability (β_N) of a nonterminal N on an automaton w given a grammar G by 8.5.

$$\beta_N(p,q) = P(w_{pq}|Npq,G) \tag{8.5}$$

Let $w_{(i,j)}$ be an automaton transition sequence between i and j consuming the symbol w. The inside probability $\beta_N(p,q)$ is the probability that a subtree rooted in N has the string yield $w_{(p,q)}$. We typically use dynamic programming via the inside algorithm to compute inside probabilities recursively. For a given nonterminal A, a PCFG G and right hand side ξ, let $P(A \to \xi|G)$ be the probability of the rule $A \to \xi$ according to G. The inside probability of a nonterminal A is defined inductively:

Base Case: For preterminal nodes A in G, $\beta_A(p,q) = P(A \to w_{p,q})|G)$.

Inductive Case: for other nonterminal nodes A in G, for binary rules of the form

$A \rightarrow B_1C_1$, $A \rightarrow B_2C_2...$, $\beta_A = \sum_{R(A) \epsilon G} \beta(B_i)\beta(C_i)$.

Renormalization by Inside Probability

Nederhof and Satta (2008) describe the computation of weighted intersection PCFG. Their method obtains the renormalized probability of a situated rule (which we situate wtih indices with (x, y),$(x_1, y_1)...$) as a product of the original probability of the unsituated rule and the inside probabilies of situated categories, according to:

$$P'(A_{(x,y)}) \rightarrow B_{(x_1,y_1)}) = \frac{r(A \rightarrow B)\beta_{B(x_1,y_1)}}{\beta_{A_{(x,y)}}} \tag{8.6}$$

$$P'(A_{(x,y)} \rightarrow B_{(x_1,y_1)}C_{(x_2,y_2)}) = \frac{r(A \rightarrow BC)\beta_{B_{(x_1,y_1)}}\beta_{C_{(x_2,y_2)}}}{\beta_{A_{(x,y)}}} \tag{8.7}$$

This requires determining the inside probability of a rule, as described in O'Donnell et al. (2009). Renormalization by inside probability of the branch reflects the true information that an incremental prefix parse contains. To condition a probabilistic context free grammar G on a finite state automata w, we need a weighted intersection G' whose categories are intersections of categories in G with state transitions in w. The probabilities on productions in G' need to somehow reflect both probabilities of rules in G and probabilities of state transitions in w.

The inside probability of a category C over a string yield w is the conditional probability of C deriving w given the grammar G and initial probabilities of productions in G. β_C moves closer to 0.0 as $w|G$ becomes less likely. Take an example where we have a grammar G with two rules, $PA \rightarrow B_1$, and $P_1A \rightarrow B_2$, where P and P_1 are the initial probabilities from training. To determine renormalized probabilities P' and P_1' on the weighted intersection of G and w, P' should be lowered relative to R' to the extent that $\beta_{B_1} \leq \beta_{B_2}$. Since the initial probabilities P and P_1 are fixed from training, the inside probabilities reflect estimated 'counts' of rules from w, which drives home the intuitive simularity between the renormalization equations in 8.6 and 8.7 and the equation for Relative Frequency Estimation, presented again in 8.8.

$$P(A \rightarrow \xi) = \frac{\sum_{i=1}^{n} f(A \rightarrow \xi; \tau_I)}{\sum_{i=1}^{n} f(A; \tau_I)} \tag{8.8}$$

This approach is distinct from pooling probability from rules that do not appear in the intersection directly over to rules that do. Given the previous example, we might set P equal to $P + P_1 = 1$ when $A \rightarrow B_2$ fails to appear in the intersection of G and w. Naive renormalization of this type can fail to propogate the information of a string event up through the grammar. When we renormalize by inside probability, we adjust the probabilities of rules whenever the inside probabilities of children non-terminals change. Intuitively, the probability of a rule is reduced whenever any path between the rule and the string yield are eliminated; inside probability propagates this information up to nonterminals higher in the parse. However, naive renormalization reduces the probability of a rule R only when all paths from R are

falsified. Naive renormalization overweights many branches because it can adjust probabilities of rules low in the tree without adjusting probabilities of ancestor branches.

8.0.7 Entropy

We can conceptualize entropy in terms of a discrete random variable. The entropy of a discrete random variable is equal to $-\Sigma_i p_i \log_2 p_i$ index. A fair coin, for example, has an entropy of $-(.5\ log\ .5) + (.5\ log\ .5)$, i.e., 1.0 bits of entropy, since each of the two possible outcomes (heads, tails) has a 0.5 probability of occurring.

Grenander (1967) demonstrates the computation of entropies of Probabilistic Context Free Grammars. For a PCFG with production rules in Chomsky normal form, let the set of production rules in G be Π, and for a given nonterminal ξ denote the set of rules with parent ξ as $\Pi(\xi))$. The entropy associated with a single rewrite of ξ is given by Equation 8.9

$$H(\xi) = -\sum r \in \Pi(\xi_i) p_r r \log_2 p_r \tag{8.9}$$

A PCFG is a random process whose outcome is a derivation, and the PCFG's total entropy is the entropy associated with derivations in Π, where each derivation is a series of rule selection events. Then the entropy of a PCFG is equal to the total

entropy of the start symbol S, where the entropy associated with one-step rewrites of ξ must inherit entropy associated with rewriting children of rules in $\Pi(\xi)$.

Grenander (1967)'s Theorem in Equation 8.10 provides a recurrence relation for determining the entropy of the start category S; each parent accrues entropy from children weighted by the probabilities of those children.

$$H(\xi_i) = h(\xi_i) + \sum r \in \Pi(\xi_i) p_r [H(\xi_{j1}) + H(\xi_{j2}) + ...] \tag{8.10}$$

The theorem also provides a closed-form solution when the probabilistic context free grammar is recursive or otherwise impractical to compute. The closed-form solution uses linear algebra to efficiently compute the entropy of the hierarchical process in two parts: the 'local' entropies of parents as simple random variables, and a fertility relation. Let $\vec{h}$ be a vector indexed by nonterminal symbols with each component given by Equation 8.11.

$$h_i = h(\xi_1) = -\sum_{r \in \Pi(\xi_i)} p_r r \log_2 p_r \tag{8.11}$$

Record the one-step fertility relation in a matrix A, labelled with non-terminals, where $A_{i,j}$ is the expected number of j is the number of i to appear in one rewriting of i. Then the vector of total entropies associated with non-terminals in G is given by Equation 8.12.

$$H_G = (I - A)^{-1}\vec{h} \tag{8.12}$$

For example, the local entropy of a category C according to a PCFG G is the entropy of a die whose sides are labeled and weighted according to one-step rewritings of C. The inversion $(I - A)^{-1}$ gives the transitive closure of the fertility relation: the expected number of j in a derivation issuing by any number of steps from i. The dot product of right hand side vector $\vec{h}$ and $(I - A)^{-1}$ gives a vector of total entropies for each non-terminal, including S.

BIBLIOGRAPHY

K. V.-S. A. K. Joshi and D. Weir. The convergence of mildly context-sensitive grammars. In S. M. Shieber and T. Wasow, editors, The Processing of Natural Language Structure.

D. Adger and J. Quer. The syntax and semantics of unselected embedded questions. Language, 77(1):107–133, 2001.

D. M. Albro. An earley-style recognition algorithm for mcfgs.

T. Alexopoulou and F. Keller. Locality, Cyclicity, and Resumption: At the Interface between the Grammar and Human Sentence Processor. Language, 83(1): 110–160, 2007.

K. V.-S. Avarind K. Joshi and D. Weir. The convergence of mildly context-sensitive grammars. In S. M. Shieber and T. Wasow, editors, The Processing of Natural Language Structure. MIT Press, 1992.

C. F. Baker, C. J. Fillmore, and J. B. Lowe. The berkeley framenet project. In Proceedings of the 17th international conference on Computational linguistics-Volume 1, pages 86–90. Association for Computational Linguistics, 1998.

A. Barss and H. Lasnik. A note on anaphora and double objects. Linguistic inquiry, 17(2):347–354, 1986.

M. Bayes and M. Price. An essay towards solving a problem in the doctrine of chances. by the late rev. mr. bayes, frs communicated by mr. price, in a letter

to john canton, amfrs. Philosophical Transactions (1683-1775), pages 370–418, 1763.

S. Berman. On the semantics and logical form of {\ it wh\/}-clauses. 1991.

R. Berwick and A. Weinberg. Parsing efficiency, computational complexity, and the evaluation of grammatical theories. Linguistic Inquiry, pages 165–191, 1982.

T. G. Bever. The cognitive basis for linguistic structure. In J. R. Hayes, editor, Cognitive development of language. Wiley, 1970.

K. Bicknell, R. Levy, and V. Demberg. Correcting the incorrect: Local coherence effects modeled with prior belief update. In Proceedings of the 35th annual meeting of the Berkeley linguistics society, pages 13–24, 2009.

S. Billot and B. Lang. The structure of shared forests in ambiguous parsing. In Proceedings of the 27th annual meeting on Association for Computational Linguistics, pages 143–151. Association for Computational Linguistics Morristown, NJ, USA, 1989.

C. Bishop et al. Pattern recognition and machine learning, volume 4. springer New York, 2006.

J. Boland, M. Tanenhaus, and S. Garnsey. Evidence for the immediate use of verb control information in sentence processing. Journal of Memory and Language, 29(4):413–432, 1990.

M. Boston, J. Hale, R. Kliegl, U. Patil, and S. Vasishth. Parsing costs as predictors of reading difficulty: An evaluation using the potsdam sentence corpus. Journal of Eye Movement Research, 2(1):1–12, 2008.

J. Bresnan. On complementizers: toward a syntactic theory of complement types. Foundations of language, 6(3):297–321, 1970.

J. Bresnan. Theory of complementation in English syntax. PhD thesis, Massachusetts Institute of Technology, 1972.

Z. Chi. Statistical properties of probabilistic context-free grammars. Computational Linguistics, 25(1):131–160, 1999.

N. Chomsky. Some concepts and consequences of the theory of government and binding, volume 6. The MIT Press, 1982.

N. Chomsky. Barriers, volume 13. MIT press, 1986.

N. Chomsky. Minimalist inquiries: The framework. In J. U. R. Martin, D. Michaels, editor, Step by Step: Essays on Minimalist Syntax in Honor of Howard Lasnik. MIT Press:Cambridge, MA, 2000.

N. Chomsky and H. Lasnik. Filters and control. Linguistic Inquiry, pages 425–504, 1977.

O. Christ. A modular and flexible architecture for an integrated corpus query system. arXiv preprint cmp-lg/9408005, 1993.

C. Clifton and L. Frazier. The use of syntactic information in filling gaps. Journal of Psycholinguistic Research, 15(3):209–224, 1986.

W. Cowart. Experimental syntax: Applying objective methods to sentence judgments. Sage Publications, 1997.

F. De Saussure. Nature of the linguistic sign. Course In General Linguistics, 1916.

A. Di Sciullo and E. Williams. On the definition of word, volume 14. Springer, 1987.

D. Dowty. Thematic proto-roles and argument selection. Language, pages 547–619, 1991.

D. Embick and A. Marantz. Architecture and Blocking. Linguistic Inquiry, 39(1): 1–53, 2007.

H. Filip, M. K. Tanenhaus, G. N. Carlson, P. D. Allopenna, and B. J. . Reduced relatives judged hard requires constraint-based analyses. In . M. P. Stevenson, S., editor, Lexical Representations in Sentence Processing, chapter Computational Psycholinguistics Series. John Benjamins, 2001.

C. J. Fillmore. The case for case. Universals in Linguistic Theory.

R. FISHER et al. Statistical methods and scientific inference. Statistical methods and scientific inference., 1956.

J. Fodor. On modularity in syntactic processing. Journal of Psycholinguistic Research, 17(2):125–168, 1988.

J. Fodor and M. Garrett. The psychological unreality of semantic representations. Linguistic Inquiry, 6(4):515–531, 1975.

R. Folli and H. Harley. On the Licensing of Causatives of Directed Motion: Waltzing Matlida All Over. Studia Linguistica, 60(2):121–155, 2006.

L. Frazier. On Comprehending Sentences: syntactic parsing strategies. PhD thesis, University of Connecticut, 1979.

L. Frazier. Sentence processing: A tutorial review. 1987.

L. Frazier, C. Clifton, and J. Randall. Filling gaps: Decision principles and structure in sentence comprehension. Cognition, 13(2):187–222, 1983.

E. Gibson. The dependency locality theory: A distance-based theory of linguistic complexity. Image, language, brain, pages 95–126, 2000.

D. Gildea and D. Jurafsky. Automatic labeling of semantic roles. Computational Linguistics, 28(3):245–288, 2002.

J. Ginzburg. Resolving questions, i. Linguistics and Philosophy, 18(5):459–527, 1995a.

J. Ginzburg. Resolving questions, ii. Linguistics and Philosophy, 18(6):567–609, 1995b.

U. Grenander. Syntax Controlled Probabilities. Brown University, 1967.

J. Grimshaw. Complement selection and the lexicon. Linguistic inquiry, 10(2): 279–326, 1979.

J. Groenendijk and M. Stokhof. Interrogative Quantifiers and Skolem functions. In K. Ehlich and H. van Riemsdijk, editors, Connectedness in Sentence, Discourse and Text, chapter Tilburg Studies in Language and Literature 4, pages 71–110. Tilburg University, Tilburg, 1983.

K. Grove. Mcfgcky: Computing information-theoretic complexity metrics on mildly context sensitive grammars. 2010.

M. Guillaumin. Conversions between mildly context sensitive grammars.

M. F. H. Seki, T. Matsumura and T. Kasami. On multiple context-free grammars. Theoretical Computer Science, 1991.

J. Hale. 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 on Language technologies, pages 1–8. Association for Computational Linguistics, 2001.

J. Hale. Grammar, uncertainty and sentence processing. PhD thesis, Johns Hopkins University, 2003a.

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

J. Hale. Uncertainity About The Rest of The Sentence. Cognitive Science, 2006.

J. Hale and E. Stabler. Strict deterministic aspects of minimalist grammars. Logical aspects of computational linguistics, pages 3–9, 2005.

K. Hale and S. Keyser. The basic elements of argument structure. MIT Working Papers in Linguistics, 32:73–118, 1998.

K. Hale and S. J. Keyser. On argument structure and the lexical expression of syntactic relations. The view from Building, 20(53-109), 1993.

M. Halle and A. Marantz. Distributed morphology and the pieces of inflection. The View from Building, 20:111–176, 1993.

M. Halle and A. Marantz. Some key features of distributed morphology. MIT working papers in linguistics, 21:275–288, 1994.

H. Harkema. Parsing Minimalist Languages. PhD thesis, University of California, Los Angeles, 2001.

H. Harley. Possession and the double object construction. Linguistic variation yearbook, 2(1):31–70, 2002.

Y. Hillel, C. Gaifman, and E. Shamir. On categorial and phrase structure grammars. Bulletin of the research council of Israel, 9, 1960.

J. Hintikka. The semantics of questions and the questions of semantics. Acta Philosophica Fennica, 28, 1976.

T. Hoekstra. Small Clause Results. Lingua, 74:101–139, 1988.

P. Hofmeister and I. Sag. Cognitive constraints and island effects. Language, 86 (2):366, 2010.

E. Jaynes and G. Bretthorst. Probability theory: The logic of science. Cambridge university press, 2003.

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

A. Joshi. Tree Adjoining Grammars: How Much Context-Sensitivity Is Required to Provide Reasonable Structural Descriptions? Natural Language Parsing: Psychological, Computational and Theoretical Perspectives, 1985.

A. K. Joshi. An introduction to tree adjoining grammars. Mathematics of language, 1:87–115, 1987.

M. Just and P. Carpenter. A theory of reading: From eye fixations to comprehension. Psychological review, 87:329–354, 1980.

L. Kallmeyer. Parsing beyond context-free grammars. Springer, 2010.

L. Karttunen. Syntax and semantics of questions. Linguistics and philosophy, 1 (1):3–44, 1977.

R. Kayne. The Antisymmetry of Syntax. MIT Press, 1994.

P. Kiparsky and J. F. Staal. Syntactic and semantic relations in pāini. Foundations of Language, 5(1):83–117, 1969.

R. Kluender and M. Kutas. Subjacency as a processing phenomenon. Language and Cognitive Processes, 8(4):573–633, 1993.

G. M. Kobele. Generating Copies: An investigation into Structural Identity in Language and Grammar. PhD thesis, UCLA, 2006.

U. Lahiri. Embedded interrogatives and predicates that embed them. PhD thesis, Massachusetts Institute of Technology, 1991.

G. Lakoff. On generative semantics. Indiana University Linguistics Club, 1969.

G. Lakoff. Toward generative semantics. Syntax and Semantics, 7:43–61, 1976.

B. Lang. Parsing incomplete sentences. In Proceedings of the 12th conference on Computational linguistics-Volume 1, pages 365–371. Association for Computational Linguistics Morristown, NJ, USA, 1988.

R. K. Larson. On the double object construction. Linguistic inquiry, 19(3):335–391, 1988.

B. Levin and M. Rappaport-Havov. Unaccusativity: at the Syntax- Lexical Semantics Interface. MIT Press, 1995.

L. Levinson. The Roots of Verbs. PhD thesis, New York University, 2007.

R. Levy. Expectation-based syntactic comprehension. Cognition, 106(3):1126–1177, 2008.

R. Levy, E. Fedorenko, and E. Gibson. The syntactic complexity of russian relative clauses. In CUNY sentence processing conference, San Diego, 2007.

R. Lewis. Falsifying serial and parallel parsing models: Empirical conundrums and an overlooked paradigm. Journal of Psycholinguistic Research, 29(2):241–248, 2000.

M. C. MacDonald. Probabilistic constraints and syntactic ambiguity resolution. Language and Cognitive Processes, (9):157–201, 1994.

C. D. Manning and H. Schütze. Foundations of statistical natural language processing. MIT press, 1999.

M. McConville. Incremental natural language understanding with combinatory categorial grammar. PhD thesis, Citeseer, 2001.

G. McKoon and R. Ratcliff. Meaning through syntax: Language comprehension and the reduced relative clause construction. Psychological Review, (110):490–525, 2003.

G. McKoon and R. Ratcliff. "Meaning Through Syntax" in sentence production and comprehension: Reply to McRae et al. (2005). Psychological Review, (112): 1032–1039, 2005.

P. Merlo and E. Ferrer. The notion of argument in PP attachment. Computational Linguistics, 32(2), 2006.

P. Merlo and S. Stevenson. Automatic verb classification based on statistical distributions of argument structure. Computational Linguistics, 27(3):373–408, 2001.

J. Michaelis. Derivational minimalism is mildly context-sensitive. Technical report, Lecture Notes in Com- puter Science, 2001.

. Miller, G. and N. Chomsky. 13. In R. D. Luce, R. R. Bush, and E. Galanter, editors, Handbook of mathematical psychology, chapter Finitary models of language users. In, pages 419–491. John iIley, 1963.

S. Munsat. Wh-complementizers. Linguistics and Philosophy, 9(2):191–217, 1986.

R. Nakanishi, K. Takada, and H. Seki. An efficient recognition algorithm for multiple context free languages. Proceedings of the Fifth Meeting on Mathematics of Language, MOL5, 1997.

M. Nederhof and G. Satta. Computing partition functions of pcfgs. Research on Language & Computation, 6(2):139–162, 2008.

E. O'Bryan. Event Structure in Language Comprehension. PhD thesis, University of Arizona, 2003.

T. J. O'Donnell, J. B. Tenenbaum, and N. D. Goodman. Fragment grammars: Exploring computation and reuse in language. 2009.

M. Palmer, D. Gildea, and P. Kingsbury. The proposition bank: An annotated corpus of semantic roles. Computational Linguistics, 31(1):71–106, 2005.

D. Perlmutter. Impersonal Passives and the Unaccusative Hypothesis. In Berkeley Linguistic Society, volume IV, pages 157–189, 1978.

D. Perlmutter and P. Postal. The 1-advancement exclusiveness law. Studies in relational grammar, 2:81–125, 1984.

D. Pesetsky. Paths and categories, unpublished phd dissertation, 1982.

D. Pesetsky. Wh-in-situ: Movement and unselective binding. The representation of (in) definiteness, 98:98–129, 1987.

D. Pesetsky. Zero syntax: Experiencers and cascades, volume 27. The MIT Press, 1996.

P. M. Postal. On raising: One rule of English grammar and its theoretical implications. mit Press Cambridge, MA, 1974.

B. Pritchett. Grammatical competence and parsing performance. University of Chicago Press, 1992.

J. Pustejovsky. The Syntax of Event Structure. The Language of Time: A Reader, 2005.

B. Roark, A. Bachrach, C. Cardenas, and C. Pallier. Deriving lexical and syntactic expectation-based measures for psycholinguistic modeling via incremental top-down parsing. In Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1, pages 324–333. Association for Computational Linguistics, 2009.

J. Ross. R. 1967. constraints on variables in syntax. Unpublished doctoral dissertation, Massachusetts Institute of Technology, 1967.

E. Sandhaus. The new york times annotated corpus. Linguistic Data Consortium, Philadelphia, 6(12):e26752, 2008.

D. Schneider and C. Phillips. Grammatical search and reanalysis. Journal of Memory and Language, 45(2):308–336, 2001.

C. Shannon. A mathematical theory of communication. the bell systems technical journal, 27: 379–423, 623–656, july, 1948.

S. Shieber, Y. Schabes, and F. Pereira. Principles and implementation of deductive parsing. The Journal of Logic Programming, 24(1-2):3–36, 1995.

J. Sprouse, M. Wagers, and C. Phillips. A test of the relation between working memory capacity and syntactic island effects. Language, 2012.

E. Stabler. Derivational minimalism. In Logical Aspects of Computational Linguistics., pages 68–95. Springer, 1997.

M. Steedman. Combinatory grammars and parasitic gaps. Natural Language & Linguistic Theory, 5(3):403–439, 1987.

M. Steedman. The Syntatcic Process. MIT Press, 2000.

S. Sternberg. The discovery of processing stages: Extensions of donders' method. Acta psychologica, 30:276–315, 1969.

S. Stevenson and P. Merlo. Lexical Structure and Parsing Complexity. Language and Cognitive Processes, 12(2), 1997.

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

L. Stowe, M. Tanenhaus, and G. Carlson. Filling gaps on-line: Use of lexical and semantic information in sentence processing. Language and Speech, 34(4): 319–340, 1991.

P. Suppes. Probabilistic grammars for natural languages. Synthese, 22(1):95–116, 1970.

M. Tanenhaus, J. Boland, S. Garnsey, and G. Carlson. Lexical structure in parsing long-distance dependencies. Journal of Psycholinguistic Research, 18(1):37–50, 1989.

R. D. Van Valin Jr and W. A. Foley. Role and reference grammar. 1980.

Z. Vendler. Verbs and times. The philosophical review, 66(2):143–160, 1957.

T. B. N. C. version 3 (BNC XML Edition). Distributed by Oxford University Computing Services on behalf of the BNC Consortium. 2007. URL `http://www.natcorp.ox.ac.uk/`.

S. Wechsler. Thematic structure. Encyclopedia of Language and Linguistics, 2nd edition. Amsterdam: Elsevier, pages 645–653, 2006.

S. Wu, A. Bachrach, C. Cardenas, and W. Schuler. Complexity metrics in an incremental right-corner parser. In Proceedings of the 48th annual meeting of the association for computational linguistics, pages 1189–1198. Association for Computational Linguistics, 2010.

J. Yun, J. Whitman, and J. Hale. Subject-object asymmetries in korean sentence comprehension. In Proceedings of the 32nd Annual Meeting of the Cognitive Science Society, volume 215272157, 2010.

M. L. Zubizaretta and E. Oh. On the Syntactic Composition of Manner and Motion, volume 48 of Linguistic Inquiries Monographs. MIT Press, 2007.