Cognitive Science 30 (2006) 643–672


# Uncertainty About the Rest of the Sentence

John Hale

*Department of Linguistics and Germanic, Slavic, Asian and African Languages, Michigan State University*

Received 26 March 2005; received in revised form 24 November 2005; accepted 5 December 2005

**Abstract**

A word-by-word human sentence processing complexity metric is presented. This metric formalizes the intuition that comprehenders have more trouble on words contributing larger amounts of information about the syntactic structure of the sentence as a whole. The formalization is in terms of the conditional entropy of grammatical continuations, given the words that have been heard so far.

To calculate the predictions of this metric, Wilson and Carroll's (1954) original entropy reduction idea is extended to infinite languages. This is demonstrated with a mildly context-sensitive language that includes relative clauses formed on a variety of grammatical relations across the Accessibility Hierarchy of Keenan and Comrie (1977).

Predictions are derived that correlate significantly with repetition accuracy results obtained in a sentence-memory experiment (Keenan & Hawkins, 1987).

*Keywords:* Linguistics; Computer science; Psychology; Syntax; Language understanding; Information; Mathematical modeling; Computer simulation; Relative clauses; Probabilistic grammars; Entropy reduction; Minimalist grammars; Accessibility hierarchy

## 1. Introduction

A *complexity metric* is a prediction about which sentences are more or less difficult to understand. In the cognitive science of language, such metrics have long held out hope of relating available experimental measures to computational theories. Miller and Chomsky (1963), for instance, expressed this hope in writing: "It might be useful, therefore, to develop measures of various sorts to be correlated with understandability" (p. 480). Examining grammatical sentences that differ in understandability, they suggested, will contribute to a more accurate picture of the human sentence processing mechanism—as opposed to grammatical competence. Indeed, since the seminal work of Yngve (1960), work in computational psycholinguistics has pursued a succession of increasingly sophisticated complexity metrics (Gibson, 1991; Kaplan, 1972; Morrill, 2000; Rohde, 2002; Stabler, 1994). Often the specification of these metrics im-

Correspondence should be addressed to John Hale, Department of Linguistics and Germanic, Slavic, Asian and African Languages, Michigan State University, East Lansing, MI 48824–1027.

plies certain constraints on the form of grammar being considered, or on the architecture of the processor.

An attractive hypothesis throughout has been that the degree of predictability of words in sentences is somehow related to understandability (Taylor, 1953) or, alternatively, production difficulty (Goldman-Eisler, 1958). Values like the predictability of a word can be obtained experimentally through procedures analogous to Shannon's (1951) guessing game. However, since the 1950s, this attractive hypothesis has lain dormant, perhaps because its integration with realistic models of linguistic structure was seen as an insurmountable challenge.

This article responds to that theoretical challenge with a complexity metric that formalizes the notion "uncertainty about the rest of the sentence" in a way that can be combined with modern proposals about syntactic structure. Section 3 introduces a way of doing this combination that involves calculating the *conditional entropy* of a grammar constrained by an initial string. Subsequent sections take up an extended psycholinguistic application in a domain—the Accessibility Hierarchy of relativizable grammatical relations (Keenan & Comrie, 1977)—where syntactic structure is centrally at stake, and a processing explanation has been sought. Before this new work is described, the existing theoretical problem is sketched in somewhat greater detail in Section 2.

## 2. A theoretical problem

Wilson and Carroll (1954) noted a tension between what they called statistical and linguistic structure. They defined an artificial language, endowed with a rudimentary phonology, morphology, and syntax, arguing that a word's informational contribution be identified with its *entropy reduction,* the downward change in average uncertainty brought about by its addition to the end of a sentence fragment. Wilson and Carroll (1954) qualified the significance of their achievement, saying:

> An entropy reduction analysis presupposes that the number of possible messages is finite, and that the probability of each of the messages is known …. Thus it appears that the entropy reduction analysis could be applied only to limited classes of natural language messages since the number of messages in nearly all languages is indefinitely large. (p. 108)

Because Wilson and Carroll's (1954) representation of linguistic structure—their grammar—is a finite list of disjunctive options, they acknowledged that it is inadequate for scaling the entropy reduction idea up to the level of sentences:

> The Whorf and Harris models of the English monosyllable could readily be used in such an analysis and lack only the conditional probabilities of passing from one state to another to be complete Markov processes. Likewise, knowledge of syntactical structure can guide us in applying entropy measures to morphemes or words and setting up appropriate Markov processes. However, it seems that existing knowledge cannot do more than provide us with guides for describing such relatively simple phenomena, leaving the more complex and less well understood aspects of linguistic structure untouched. (p. 103)

The "more complex and less well understood aspects of linguistic structure" can be read here as a blanket characterization of all aspects of natural language syntax extending beyond the generative capacity of nonlooping finite-state automata. Indeed, Chomsky's (1956) argument that English not be considered any sort of finite-state language highlights the theoretical problem: integrating a natural formalization of cognitive effort (in terms of information-theoretic work) with a realistic conception of linguistic structure.

However, with Wilson and Carroll's (1954) proposal reinterpreted as actually incorporating a linguistic theory—albeit one that is too weak—the question arises what their idea would look like with a different, more adequate grammar. In particular, can the entropy reduction idea be combined with one of the mildly context-sensitive grammars? Members of this class are widely agreed to be expressive enough to accommodate reasonable structures for natural language sentences while still ruling out some conceivable alternatives (Frank, 2004; Joshi, Vijay-Shanker, & Weir, 1991).

Section 3 answers this question in the affirmative, showing how the entropy reduction idea can be extended to mildly context-sensitive languages by applying two classical ideas in (probabilistic) formal language theory: Grenander's (1967) closed-form solution for the entropy of a nonterminal in a probabilistic grammar, and Lang's (1974, 1988) insight that an intermediate parser state is itself a specification of a grammar. Sections 4 through 10 assert the feasibility of this extension by examining the implications of two alternative relative clauses analyses for a proposed linguistic universal. In different ways, these two analyses make essential use of the nonconcatenative devices needed to describe unbounded dependency in natural language. They therefore exercise the greater flexibility provided by the abstract perspective on entropy reduction presented in Section 3. With the theoretical problem overcome, a more cognitive perspective on intermediate parser states emerges, in which a theory of psychological effort (as formalized by information-theoretic uncertainty) is parameterized by explicit, changeable linguistic assumptions.

## 3. Entropy reduction

The idea of the entropy reduction of a word is that uncertainty about grammatical continuations fluctuates as new words come in. Entropy, as a fundamental concept in information theory (Shannon, 1948), formalizes uncertainty about specified alternatives. The notion of grammatical continuation can be similarly grounded in the concept of derivations consistent with some words already seen.

Probabilistic context-free grammars (PCFGs)[1] can play this definitional role. The derivations consistent with words already seen can be examined directly in a very small PCFG like the one in Fig. 1. This tiny grammar generates two sentences. There are exactly two derivations, each with probability one-half.

$$\begin{array}{llll} 0.5 & \mathrm{S} & \rightarrow & \text{john loves Mary} \\ 0.5 & \mathrm{S} & \rightarrow & \text{john sleeps} \end{array}$$

Fig. 1. A very simple PCFG.

On hearing the first word of a sentence in the language of this grammar, no information is conveyed to the hearer. Because all derivations are compatible with the prefix string "john … " it is only the second word that eliminates one of the derivations, conveying a single bit. Said another way, finding out that the second word is "loves" rather than "sleeps" reduces the entropy of the start symbol S from 1 bit to 0 bits. On this view, nonterminals like S are random variables that have as outcomes the right-hand sides of rules that rewrite them.

In Fig. 1 it is easy to see what the set of alternative derivations is; in principle its members could be written out. Such enumeration is not effectively possible in a recursive grammar like the one in Fig. 2.

| | | | |
|---|---|---|---|
| 0.5 | S | → | john thinks CP |
| 0.5 | S | → | john sleeps |
| 1.0 | CP | → | that S |

Fig. 2. Recursive PCFG.

The grammar in Fig. 2 defines an infinite number of derivations, the probability of which tails off with their size so that the sum of the probabilities assigned to the language is exactly 1.0.

| word # | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | ⋯ |
|---|---|---|---|---|---|---|---|---|---|
| | john | sleeps | | | | | | | |
| | john | thinks | that | john | sleeps | | | | |
| | john | thinks | that | john | thinks | that | john | sleeps | |
| | | | | | ⋮ | | | | |

Fig. 3. Sentences generated by recursive PCFG.

On this grammar two bits are conveyed if "sleeps" is the second—or fifth, eight, eleventh, and so on—word. (See Fig. 3.) Even though there are an infinite number of derivations licensed by this grammar, it is possible to compute the entropy of the start symbol by applying a theorem due to Grenander (1967).

### *3.1. Entropy of nonterminals in a PCFG*

Grenander's theorem is a recurrence relation that gives the entropy of each nonterminal in a PCFG $G$ as the sum of two terms. Let the set of production rules in $G$ be $\Pi$ and the subset rewriting the $i$th nonterminal $\xi_i$ be $\prod(\xi_i)$. Denote by $p_r$ the probability of a rule $r$.

Then

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

$$H(\xi_i) = h(\xi_i) + \sum_{r\in\Pi(\xi_i)} p_r\left[H(\xi_{j1}) + H(\xi_{j2}) + \cdots\right]$$

The first term, lowercase $h$, is simply the definition of entropy for a discrete random variable. The second term, uppercase $H$, is the recurrence. It expresses the intuition that derivational uncertainty is propagated from children (the $\xi_{j1}$, $\xi_{j2}$, …) to parents ($\xi_i$).

For PCFGs that define a probability distribution on the generated language, the solution to this recurrence can be written as a matrix equation where $I$ is the identity matrix, $\vec{h}$ the vector of the $h_i$ and $A$ is a matrix whose $(i, j)$th component gives the expected number[2] of nonterminals of type $j$ resulting from nonterminals of type $i$.

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

(Grenander, 1967, p. 19).

Continuing with the grammar of Fig. 2

$$\begin{aligned} h_{\mathrm{S}} &= -0.5\times\log_2 0.5 - 0.5\times\log_2 0.5 &&= 1 \\ h_{\mathrm{CP}} &= -1.0\times\log_2 1.0 &&= 0 \end{aligned}$$

there is one bit of uncertainty about the choice of S rule, and no uncertainty about the choice of CP rule, so the vector of single rule entropies, $h$ is <1, 0>. As far as the expected number of children of each type, the expectation for S deriving any CPs is only one half. On the other hand, CP will definitely rewrite as S. No other children can be derived, leaving $A$ entirely defined.

$$A = \begin{pmatrix} 0 & \frac{1}{2} \\ 1 & 0 \end{pmatrix}$$

The recurrence that Equation 1 solves is

$$\begin{aligned} H &= \vec{h} + AH \\ \vec{h} &= H - AH = (I - A)H \\ H &= (I - A)^{-1}\vec{h} \end{aligned}$$

In the case of grammar 2,

$$H = \left[\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} - \begin{pmatrix} 0 & \frac{1}{2} \\ 1 & 0 \end{pmatrix}\right]^{-1} \begin{pmatrix} 1 \\ 0 \end{pmatrix}$$

$$H = \begin{pmatrix} 2 \\ 2 \end{pmatrix}$$

the calculation delivers the answer that the entropy of both S and CP in the grammar of Fig. 2 is 2 bits.

### 3.2. Incomplete sentences

Grenander's theorem supplies the entropy for any PCFG nonterminal in one step by inverting a matrix. To determine the contribution of a particular word, one would like to be able to look at the change in uncertainty about compatible derivations as a given prefix string is lengthened. When the set of such derivations is finite, it can be expressed as a list. In the case of a recursive grammar such as the one in Fig. 2 that has an infinity of compatible derivations, some other representation is necessary.

Lang and Billot (Lang, 1974; Billot & Lang, 1989) pointed to what this other representation might be. They show how parsing, in general, can be viewed as the intersection of a grammar with a finite-state language. If the grammar is context free, then the intersection with a finite-state language will itself be context free (Bar-Hillel, Gaifman, & Shamir, 1960). This perspective readily accommodates the view of incomplete sentences as finite-state languages with members that all have the same initial *n* words but optionally continue on with all possible words of the terminal vocabulary, for all possible lengths. Using the question mark to symbolize any vocabulary word, Fig. 4 illustrates a finite-state representation of the initial three words of a sentence.

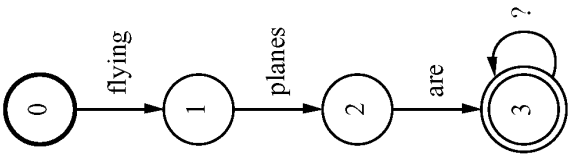



Fig. 4. Finite state machine specifying a prefix.

The intersection of an automaton like the one in Fig. 4 with a grammar *G* is another grammar specifying all and only the *G*-licensed continuations of the string "flying planes are." This result is often encoded in a data structure called a chart.

### 3.3. Chart parser states as grammars

The status of context-free grammars as the result of intersecting a finite state input with a context-free grammar can be appreciated by looking at chart parsing (Kay, 1986). The chart in chart parsing is typically a two-dimensional array, with cells that record sets of nonterminals in some way or another. As parsing happens, the array becomes populated with information about the presence of constituents, for example, that there is a noun phrase (NP) spanning positions 2 through 5. This would be indicated by adding NP to the set of nonterminals in the (2, 5) cell of the chart. In a recognizer, all that matters is the presence or nonpresence of the start symbol in the cell with left and right positions that correspond to the beginning and end of the input sentence. In a parser, derivations must be recovered. This is typically done by augmenting chart cells with back pointers that record the addresses of any child constituents. The key point is that a pair of back pointers picking out, for instance, Determiner in cell (2, 3) and Noun in cell (3, 5) define a kind of grammar rule (Fig. 5).

$$\mathrm{NP}_{(2,5)} \rightarrow \mathrm{D}_{(2,3)} \quad \mathrm{N}_{(3,5)}$$

Fig. 5. "Situated" grammar rule.

This kind of grammar rule refers not just to nonterminals in general, but to nonterminals situated in the string being parsed. Crucially, the chart keeps only single entries for nonterminals participating in derivational loops. For instance, if an NP can be a bare noun and a bare noun can be an NP, then a chart parser that has found an $NP_{(2,5)}$ would also insert $N_{(2,5)}$ in the same cell. Back pointers would be set up both from NP to N and from N to NP. In this way, back pointers record the possibility of recursion without falling into an infinite loop evaluating such recursion.

### 3.4. Generalization to other formalisms

The symbol names and back pointers in a parser's chart thus specify the intersection of the parser's grammar with a given finite-state input. Just as chart parsing extends to more expressive formalisms, so too does the representation of derivations as chart grammars in more expressive formalisms like Synchronous Tree-Adjoining Grammars (Shieber & Johnson, 1993) and Definite Clause Grammars (van Noord, 2001). Chart entries in mildly context-sensitive grammars, such as the Minimalist Grammars (MGs) used in Section 5 will in general have more than two indexes into their input, whether it be a string or a more complicated finite-state language. However these are just part of longer nonterminal names that are completely analogous to the situated grammar rules obtained in context-free parsing. The string manipulation functions ordinarily associated with each branch (Michaelis, 1998; Seki, Matsumura, Fujii, & Kasami, 1991; Weir, 1988) do not even need to be represented if their choice is unambiguous given the branch, as it is in MGs (Hale & Stabler, 2005). If a rule is reachable from a start symbol spanning the entire string, then it defines one branch of a possible derivation tree for the given input.

### 3.5. Conditional entropy

In light of Lang and Billot, the set of grammatical derivations of a string starting with $w = w_0 w_1 \ldots w_n$ is well-defined and can be computed (by chart parsing) for any formalism that is closed under intersection with the finite-state languages. The set of compatible continuations is intensionally represented by a grammar generating derivation trees with labels that are annotated with state name indexes as in Fig. 5. This "situatedness" persists even down to the terminals: the grammar of continuations defines a language of situated lexical entries.

This grammar can be taken as the categorical part of a related probabilistic grammar, weighted according to a priori expectations, with entropy that can be calculated using Grenander's theorem. Altogether, the extension of Wilson and Carroll's (1954) idea to infinite state grammars involves the five steps listed in Fig. 6.

1. viewing initial substring $w$ as a finite state language $w(?)^*$
2. intersecting $w(?)^*$ with $L(G)$
3. to obtain a chart `Ch` of parses generating $w$ (Lang, 1988)
4. which can be viewed as a probabilistic grammar `pcfg_of`$(G, \texttt{Ch})$
5. the entropy of whose start symbol can be calculated (Grenander, 1967)

Fig. 6. How to calculate the conditional entropy of a word in a sentence.

The recipe in Fig. 6 generalizes the method of Hale (2003b), which is founded on top-down parsing of non-left-recursive, context-free grammars. As in that work, the sentence-level difficulty predictions can be derived by adding up individual word difficulty predictions.

### *3.6. The entropy reduction hypothesis*

Such extension allows Wilson and Carroll's (1954) idea to be stated more generally. Let `Start` be the start symbol of a probabilistic grammar $G$, and $w_0 \ldots w_i$ be an initial substring of a sentence in the language of $G$. Then, abbreviating the conditional entropy of an $i$-word string on a grammar $G$ (or $H_G(\text{Start}|w_0 \ldots w_i)$) as $H_i$, define the disambiguation work $ER$ done at a word $w_i$ as

$$ER(w_i) = \max(0,\ H_{i-1} - H_i) \tag{2}$$

Definition 2 uses the downward change in conditional entropy[3] to assert that "work" must result in "progress." When a word takes a sentence processor from a state of relative indifference (high $H_{i-1}$) to one of more localized suspicion (low $H_i$) comparatively more work has been done. Perhaps counterintuitively, if a processor hears a word that leaves it in a more uncertain state than before, no progress disambiguating the string has occurred, analogous to pushing against a heavy boulder on an incline, which nonetheless drives the pusher backward.

> Entropy reduction hypothesis (ERH): *ER* is positively related to human sentence processing difficulty.

The ERH might be glossed as saying that a person's processing difficulty at a word in a sentence is proportional to the number of bits signaled to the person by that word with respect to a probabilistic grammar the person knows. It is a complexity metric whose predictions that are deducible using the method in Fig. 6. Indeed, because the method is compatible with so many probabilistic grammars, Section 4 begins an extended example of its use with relative clauses, a domain in which the controversial details of grammatical structure have been argued to take on a universal significance.

## 4. The Accessibility Hierarchy

The Accessibility Hierarchy (AH) is a cross-linguistic generalization about relative clause formation discovered by Keenan and Comrie (1977). Relative clauses have long held particular interest for sentence processing theorists as examples of long-distance dependency, as in Example 1.

(1) a. the father explained the answer to the boy
b. * the boy who the father explained the answer to the boy was honest
c. the boy who the father explained the answer to was honest

Sentence 1a is perfectly acceptable on its own. However, when embedded in the context *the boy who … was honest,* the result 1b is unacceptable. Acceptability can be restored by removing the inner copy of *the boy*, arriving at 1c. It is as if the sentence-initial NP *the boy* de-

SUBJECT ⊃ DIR. OBJECT ⊃ INDIR. OBJECT ⊃ OBLIQUE ⊃ GENITIVE ⊃ OCOMP

Fig. 7. The Accessibility Hierarchy of relativizable grammatical relations.

pends on the nonpresence of its subsequent repetition to guarantee the acceptability of the whole sentence. The dependency is long distance because there may be an arbitrary amount of embedding between, for example, *the boy* and the position immediately following the preposition *to*. Whatever part of the human processor considers *the boy* a legitimate perceptual option following *the answer* in 1a but not in 1b might as well be identified as a relativization rule of English.

The AH, then, is a scale of grammatical relations borne by the dependent element prior to relativization. For instance, 1c is an example of relativization from Indirect Object because *the boy* is the indirect object in 1a. Much work (Bever, 1970; Gibson, 1998; Wanner & Maratsos, 1978) has focused on relativization from Subject and Object, but of course, there are other grammatical relations, some of which support relativization in particular languages.

The cross-linguistic AH generalization is that every language occupies a position on the scale (shown in Fig. 7) and that relativization is possible from any grammatical relation to the left of that language-particular point.

The AH[4] figures in a variety of modern syntactic theories that have been influenced by relational grammar (Perlmutter & Postal, 1974). In Head-Driven Phrase Structure Grammar (Pollard & Sag, 1994) the Hierarchy corresponds to the order of elements on the SUBCAT list, and interacts with other principles in explanations of binding facts. The Hierarchy also figures in Lexical-Functional Grammar (Bresnan, 1982) where it is known as Syntactic Rank or the Relational Hierarchy.

### *4.1. The Keenan and Hawkins experiment*

Keenan and Comrie (1977) speculated that their typological generalization might have a basis in performance factors. This idea was supported by the results of a psycholinguistic experiment done in 1974 that were not published until much later (Keenan & Hawkins, 1987).

Seeking some processing basis for the AH, Keenan and Hawkins conducted a study that examined people's ability to comprehend, remember, and produce sentences involving relative clauses from various points on the AH. They constructed four stimulus sentences exhibiting relativization on each grammatical relation[5] using relatively frequent words—at least 50 per million in materials sampled by Thorndike and Lorge (1968) in the 1930s.

Both adults and children participated in the experiment; all were native British English speakers. Participants heard each stimulus sentence read out loud on a tape. Half a second after the last word, adult participants heard the names of eight digits. They were then expected to write down this digit sequence. Having completed the digit-memory interference task, participants were finally asked to demonstrate their memory for the original stimulus sentence by writing it. These responses were coded for accuracy according to a point scheme.

> Responses were graded on a 2, 1, 0 basis with 2 being the best and 0 the worst. A grade of 2 was assigned for essentially perfect repetition, allowing only a grammatically permissible change in rela-

tive pronoun (e.g., "that" for "which") and a grammatically permissible deletion of a relative pronoun.

A grade of 0 was assigned if the response did not include a relative clause where the head has the same function as the stimulus … .

In other cases errors were regarded as minor and the response assigned value 1, e.g., tense change, verb particle omission, omission or incorrect addition of a noun modifier, lexical substitution with meaning retained, substitution of one proper noun for another, and incorrect recall of either the initial frame or the final transitive verb phrase. (Keenan & Hawkins, 1987, p. 68)

Under this coding scheme, response accuracy drops as the grammatical relation relativized on becomes more rare[6] in the world's languages (Fig. 8). These scores are then summed across adult and child participants; overall children followed the AH more closely than adults.[7]

| Grammatical Relation: | SU | DO | IO | OBL | GenS | GenO |
|---|---|---|---|---|---|---|
| Repetition Accuracy: | 406 | 364 | 342 | 279 | 167 | 171 |
| errors (= $\text{R.A.}_{\max}$ − R.A.) | 234 | 276 | 298 | 361 | 471 | 469 |

Fig. 8. Results from Keenan and Hawkins (1987).

Keenan and Hawkins (1987) conclude that "the AH is reflected in the psychological processing of relative clauses in that repetition errors increase as the position of the relative clause on the hierarchy decreases" (p. 83). However they were careful to say, "It remains unexplained just why RCs should be more difficult to comprehend-produce as they are formed on positions lower on the AH" (p. 82).

The ERH, if correct, would offer just such an explanation. If a person's difficulty on each word of a sentence is related to the amount of derivational information signaled by that word, then the total difficulty reading a sentence should be the sum of the difficulty on each word. The explanation would be that sentence understanders in general must do information-processing work on the scale of the entropy reduction brought about by each word, and that this work, on a per-sentence level, increases with the AH. This explanation would also be a cognitive one, in referring to uncertainty associated with time-indexed internal states of the comprehender. Indeed, in specifying these states it would be odd to use anything other than the same linguistic grammar that describes speakers' intuitions about entire sentences. For this reason, Section 5 goes into two alternative ways of analyzing the syntactic structure of relative clauses, both of which are adequate for Keenan and Hawkins's stimuli. Average repetition accuracy results are available for each grammatical relation tested; the stimuli themselves are listed in their entirety in Fig. 9.

## 5. Relative clauses analyses

The structure of relative clauses remains a contentious issue in linguistic theory (chronologically: Sag, 1997; Borsley, 1997; Bianchi, 2000; Aoun & Li, 2003). Within transformational grammar, special care has been taken to preserve a theoretical connection between full sentences like 1a and relative clauses like 1c, repeated here as 2a and 3a. Because both syntactic constructions are subject to the same kind of cooccurrence restrictions, they are held to sit in a

| *subject* | *oblique* |
|---|---|
| they had forgotten that the boy who told the story was so young | they had forgotten that the box which Pat brought with apples in was lost |
| the fact that the girl who paid for the tickets is very poor doesn't matter | the fact that the girl who Sue wrote the story with is proud doesn't matter |
| I know that the girl who got the right answer is clever | I know that the ship which my uncle took Joe on was interesting |
| he remembered that the man who sold the house left the town | he remembered that the food which Chris paid the bill for was cheap |
| *direct object* | *genitive subject* |
| they had forgotten that the letter which Dick wrote yesterday was so long | they had forgotten that the girl whose friend bought the cake was waiting |
| the fact that the cat which David showed to the man likes eggs is strange | the fact that the boy whose brother tells lies is always honest surprised us |
| I know that the dog which Penny bought today is very gentle | I know that the boy whose father sold the dog is very sad |
| he remembered that the sweets which David gave Sally were a treat | he remembered that the girl whose mother sent the clothes came too late |
| *indirect object* | *genitive object* |
| they had forgotten that the man who Ann gave the present to was old | they had forgotten that the man whose house Patrick bought was so ill |
| the fact that the boy who Paul sold the book to hates reading is strange | the fact that the sailor whose ship Jim took had one leg is important |
| I know that the man who Stephen explained the accident to is kind | I know that the woman whose car Jenny sold was very angry |
| he remembered that the dog which Mary taught the trick to was clever | he remembered that the girl whose picture Clare showed us was pretty |

Fig. 9. Stimuli from Hawkins and Keenan (1974/1987).

paradigmatic relation. Examples 2 and 3 illustrate how both constructions share the requirement for a concrete noun, but enforce it in different positions.

(2) a. the father explained the answer to the boy
    b. *the father explained the answer to justice
(3) a. the boy who the father explained the answer to
    b. *justice the father explained the answer to

### *5.1. The adjunction analysis*

Perhaps the most standard analysis of relative clauses in generative grammar holds that the WH-movement rule (which is also responsible for question formation, clefts, comparatives, topicalization, etc.) transformationally relates full sentences and relative clauses (Chomsky, 1977). This analysis derives an example like 3a from an underlying string of the form *the father explained the answer to who*. WH-movement rearranges the tree structure of this underlying form, and the result is permitted as an adjoined postmodifier of an NP.

The structural description in Fig. 10 illustrates both aspects of this analysis.[8] In this picture, the WH-word *who* is categorized as a determiner phrase (or dP) just like *the boy*. The zero in-

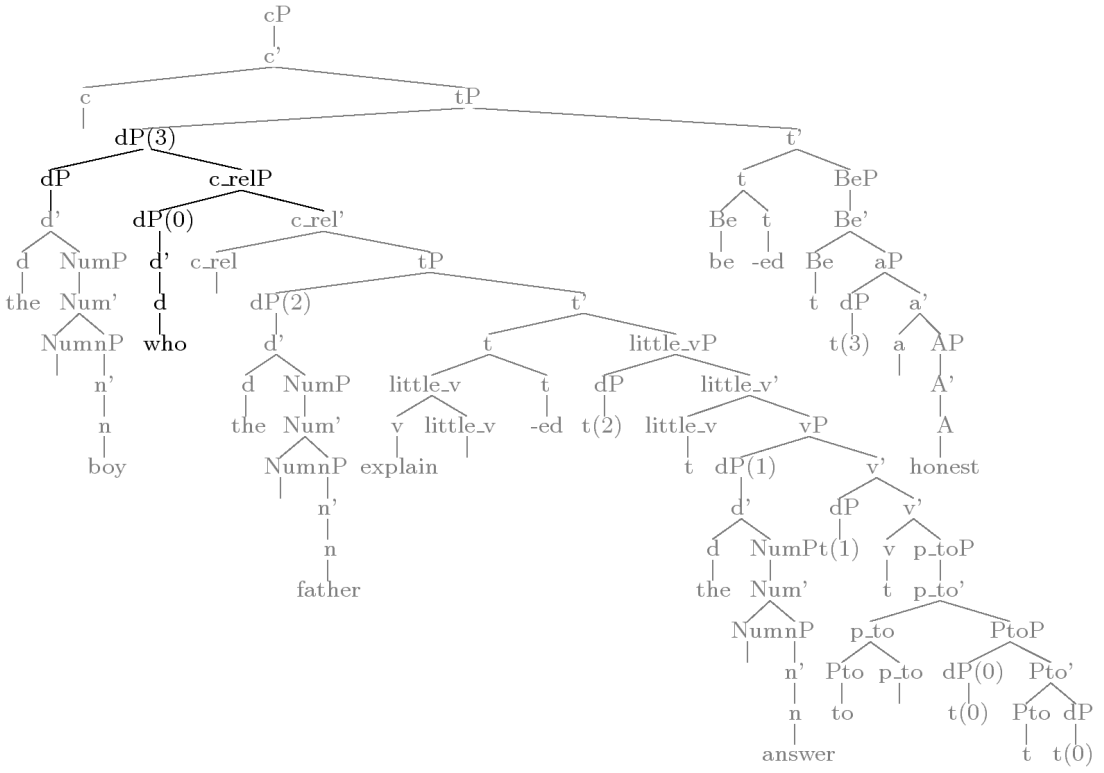



Fig. 10. Adjunction analysis.

dex on *who*'s dP node is shared by various inaudible traces (notated *t*(0)) that mark the tree locations where movement originated. Here, [ $_{dp(0)}$ *who*] was moved from a position in between the words *to* and *be*, where it was the complement of *to*: the impossibility of an incompatible complement such as *justice* in that position derives the unacceptability of both 2b and 3b. The theory postulates that WH-movement deposits *who* at the leftmost edge of the constituent *who the father explained the answer to,* a complementizer phrase, or c_relP. The reason the analysis is called the adjunction analysis is because this relative CP is adjoined, like other postmodifiers, to [*dP the boy*] in a configuration described by the phrase structure rule dP → dP c_relP. This adjunction configuration is indicated more darkly in Fig. 10.

### *5.2. The promotion analysis*

An alternative promotion analysis supposes that it is not just the WH-word that moves in relativization. The head noun *boy* also originates as a complement of the preposition, only to be promoted to the left edge of the phrase by a transformation. One argument for this analysis uses the properties of idiomatic expressions. For instance, Brame (1976) noted that a word such as *headway* is restricted to the fixed phrase *make headway* as in Example 4.

(4) We made headway.

The same property holds for other idioms such as *keep careful track of, pay lip service to, take umbrage at,* and so on. Without the special verb *make,* the idiom becomes unacceptable as in Example 5.

(5) *(The) headway was satisfactory.

Example 5 is presumably unacceptable because of the same sort of cooccurrence restriction as in Examples 2 and 3. However, in a relative clause, *headway* can occur disconnected from *make* as in Example 6.

(6) The headway that we made was satisfactory.

Brame (1967, quoted in Schachter, 1973) argued that Example 6 is grammatical because, at some stage of the derivation where the relevant cooccurrence restriction was enforced, *headway* indeed was a complement of *make*. The analysis is that *headway* has been transformationally promoted from a position adjacent to *make*, where it would be said to have been "base-generated," to its surface position between the determiner and the complementizer. This is the same pattern of argument from cooccurrence restrictions used to establish a transformational relation between active and passive sentences.

The promotion analysis was revived by Kayne (1994) for reasons having to do with Kayne's general theory of phrase structure. In this theory, adjunction is banned, so Kayne rejects the earlier analysis of relative clauses as adjoined noun modifiers, proposing instead that the underlying form of a sentence like Example 3a is akin to Example 7.

(7) the father explained the answer to $[_{\text{dP[+wh]}}$ who boy$_{\text{[+f]}}]$

On this analysis, *who* is a determiner that, in the underlying form, occupies a dP along with *boy.* Movement applies twice to discharge formal features occuring on particular words and phrases. One movement, triggered by a property all WH-words share (subscripted +wh), transports the determiner phrase [$_{dP}$ *who boy* ] to the left edge of the relative clause. Another movement, triggered by another lexical feature (+f), makes *boy* more prominent, leaving the wh-determiner *who* and the common noun *boy* in their attested linear order. A particular lexical entry for *the* then allows it to take a relative CP as a complement.

As indicated in Fig. 11, no adjunction is used in this derivation, and, unconventionally, the leftmost *the* and *boy* do not share an exclusive common constituent.

Support from the promotion analysis is bolstered by analyses of relative clauses in various languages, many of which are detailed in the edited collection of Alexiadou, Law, Meinunger, and Wilder (2000).

### *5.3. Minimalist grammars of relativization*

The syntactic analyses of relative clauses in Sections 5.1 and 5.2 are both transformational in nature. They derive the surface structure of an example like Example 3a from an underlying form in which [*who*] or [*who boy*] is an object of the embedded verb *explained*. Formally, the relativization rules these theories propose relate trees to trees.

To use either analysis in a general processing hypothesis such as the ERH, it is necessary to concretely specify these transformations, and to apply them in defining sets of possible gram-

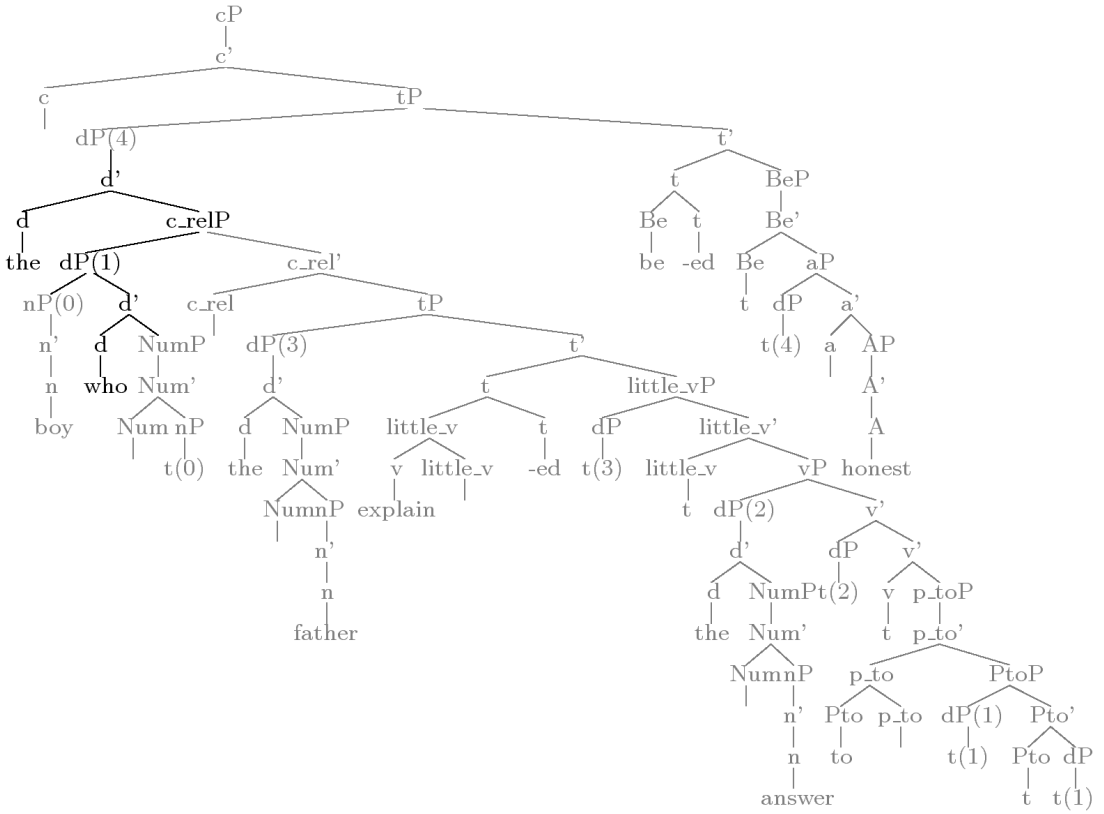


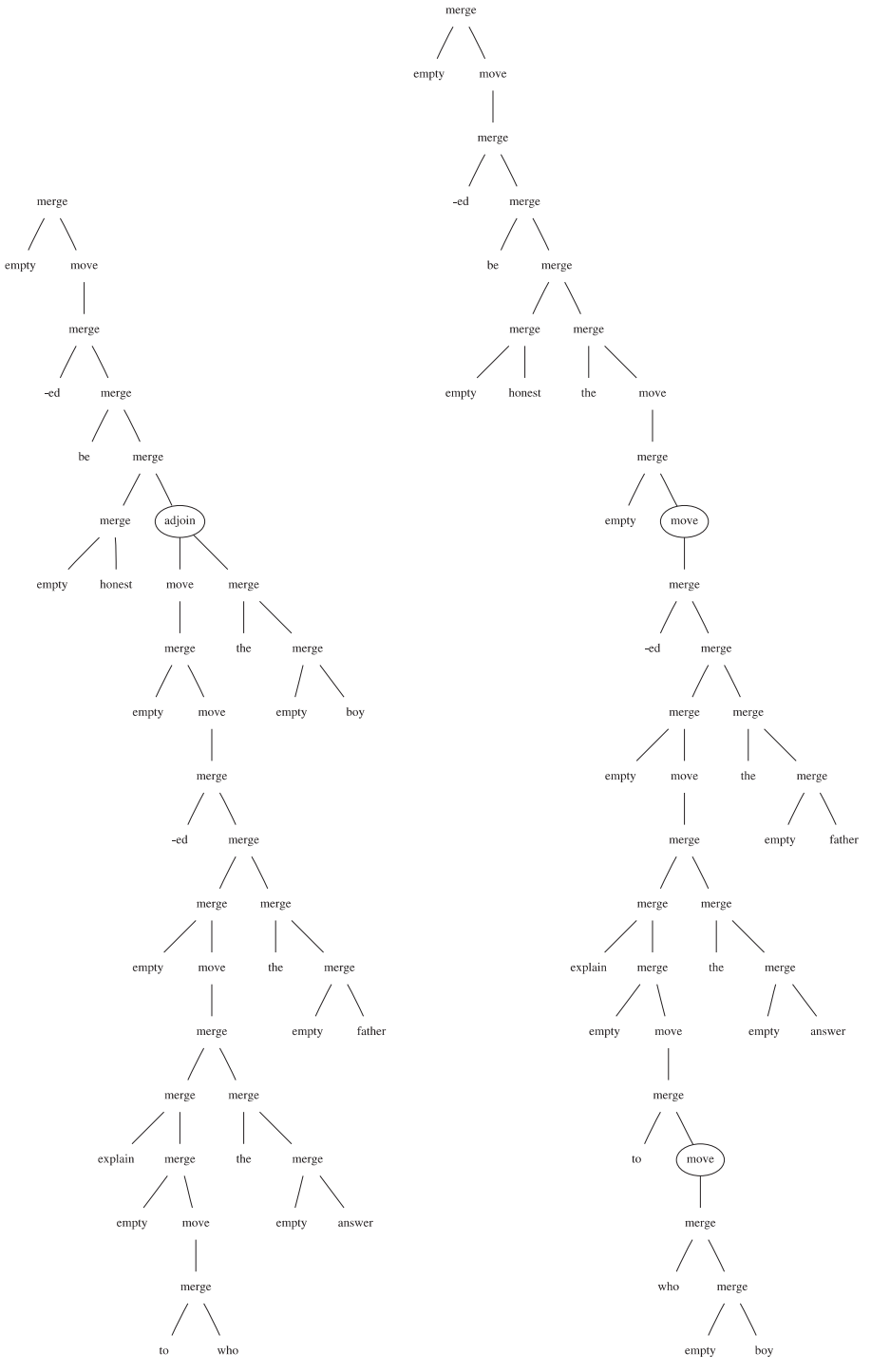

Fig. 11. Kaynian promotion analysis.

matical continuations. It would be difficult to calculate such consequences if these relativization rules were formalized as in classical transformational grammar (Peters & Ritchie, 1973). Both treatments, however, find a natural expression as MGs (Stabler, 1997), a mildly context-sensitive formalism (Joshi, 1985).

Mildly context-sensitive grammars can derive natural syntactic descriptions of sentences exhibiting crossing dependencies (Kroch & Joshi, 1985; Stabler, 2004) and other nonconcatenative phenomena. Both the adjunction and promotion analyses assert this kind of description in claiming that *who* or *who boy* appears in one position on the surface but is subject to grammatical constraints in another position.

MGs in particular are well-suited for expressing transformational analyses of phenomena like relativization. Directly formalizing certain themes of Chomsky's (1995) Minimalist Program, MGs have concatenative structure-building rules "merge" and "adjoin" that combine phrases into larger phrases. However MGs also have a nonconcatenative rule called "move." This singulary transformation can rearrange words that have already been derived, thus implementing the transformational intuition that surface structures are derived from other structures that have the same form.

Fig. 12 shows the history of merge, move, or adjoin operations that derive the surface structures 10 and 11 on particular MGs. Note that leaves of these derivations do not come in the




(a) Derivation tree on adjunction grammar (b) Derivation tree on promotion grammar

Fig. 12. Minimalist grammar derivation trees for "the boy who the father explained the answer to was honest."

same order as the words in the string Examples 1c and 3a. This is because the move rule is nonconcatenative: It can rearrange words that have already been derived, for instance, transporting *who* out of a merge-built phrase with *boy* up to the leftmost edge of a relative CP. The circled nodes highlight the distinction between adjoining a relative clause in which only the WH word has moved in Fig. 12a and merging a complement that includes a WH-determiner phrase, out of which the relative head *boy* has already moved. Such evacuation from within a moved phrase is implemented[9] by the two circled move operations in Fig. 12b.

As in Tree Adjoining Grammars (Joshi, Levy, & Takahashi, 1975), derivation trees encode everything there is to know about an MG derivation. They are the central element in the proof of MGs' place in the Chomsky Hierarchy (Michaelis, 1998) and can be parsed in a variety of orders (Harkema, 2001). Indeed, because MG and other mildly context-sensitive grammars' derivations are tree-shaped, they can be given the same branching process interpretation that underlies purely concatenative probabilistic grammars. The parameters of such a probabilistic model can be set using any PCFG estimation method, as detailed in the next section.

## 6. Procedure

Seeking an explanation for Keenan and Hawkins's (1987) psycholinguistically supported AH, two MGs were constructed to cover their experimental stimuli (listed in Fig. 9). One grammar adopts Kayne's (1994) version of the promotion analysis, whereas the other uses the more standard adjunction analysis. These two grammars (discussed in detail in Hale, 2003a) are close variants of one another, and are both relatively small, comprising about 50 types of lexical entries.

To estimate probabilistic versions of these grammars, Hawkins and Keenan's 24 stimuli were viewed as a micro-treebank from which context-free derivation-rule probabilities could be read off using the usual relative frequency estimation technique (Chi, 1999).

There are, by design, exactly four examples of relativization from each grammatical relation, so derivation tree branches from each of the six levels of treatment were weighted according to corpus data from the eight subsets of the Brown corpus (Kučera & Francis, 1967) that are syntactically annotated in the third release of the Penn Treebank (Marcus et al., 1994).

| count | grammatical relation | definition |
|---|---|---|
| 1430 | subject | co-indexed trace is the first daughter of S |
| 929 | direct object | co-indexed trace is immediately following sister of a V-node |
| 167 | indirect object | co-indexed trace is part of a PP not annotated as benefactive, locative, manner, purpose, temporal or directional |
| 41 | oblique | co-indexed trace is part of a benefactive, locative, manner, purpose, temporal or directional PP |
| 34 | genitive subject | WH word is *whose* and co-indexed trace is first daughter of S |
| 4 | genitive direct object | WH word is *whose* and co-indexed trace is immediately following sister of a V-node |

Fig. 13. Counts from Brown portion of Penn Treebank III.

| | | | | |
|---|---|---|---|---|
| 1.0000 | +case +dat v,-case -wh_rel,-dat | → | =d +case +dat v,-dat | d -case -wh_rel |
| 1.0000 | +case +dat v,-case,-dat,-wh_rel | → | =d +case +dat v,-dat,-wh_rel | ::d -case |
| 1.0000 | +case =d a,-case | → | ::=d +case =d a | d -case |
| 1.0000 | +case Pfor,-case | → | ::=d +case Pfor | d -case |
| 1.0000 | +case Pfor,-case -wh_rel | → | ::=d +case Pfor | d -case -wh_rel |
| 1.0000 | +case Pin,-case -wh_rel | → | ::=d +case Pin | d -case -wh_rel |
| 1.0000 | +case Pon,-case -wh_rel | → | ::=d +case Pon | d -case -wh_rel |
| 1.0000 | +case Pto -dat,-case | → | ::=d +case Pto -dat | ::d -case |
| 1.0000 | +case Pto -dat,-case -wh_rel | → | ::=d +case Pto -dat | d -case -wh_rel |
| 1.0000 | +case Pto,-case | → | ::=d +case Pto | d -case |
| 1.0000 | +case Pto,-case -wh_rel | → | ::=d +case Pto | d -case -wh_rel |
| 1.0000 | +case Pwith,-case -wh_rel | → | ::=d +case Pwith | d -case -wh_rel |
| 0.1147 | +case t,-case | → | ::=¿Have +case t | Have,-case |
| 0.4355 | +case t,-case | → | ::=¿Be +case t | Be,-case |
| 0.0648 | +case t,-case | → | ::=little_v +case t | little_v,-case |
| 0.3850 | +case t,-case | → | ::=¿little_v +case t | little_v,-case |
| 1.0000 | +case t,-case -wh_rel | → | ::=¿little_v +case t | little_v,-case -wh_rel |
| 1.0000 | +case t,-case,-wh_rel | → | ::=¿little_v +case t | little_v,-case,-wh_rel |
| 0.8316 | +case v,-case | → | ::=d +case v | d -case |
| 0.1684 | +case v,-case | → | ::=d +case v | ::d -case |
| 0.6681 | +case v,-case -wh_rel | → | ::=d +case v | d -case -wh_rel |
| 0.3319 | +case v,-case -wh_rel | → | =d +case v | d -case -wh_rel |
| 1.0000 | +case v,-case,-wh_rel | → | =d +case v,-wh_rel | d -case |
| 1.0000 | +dat v,-dat,-wh_rel | → | +case +dat v,-case,-dat,-wh_rel | |
| 1.0000 | +dat v,-wh_rel,-dat | → | +case +dat v,-case -wh_rel,-dat | |
| 1.0000 | +f d -case -wh_rel,-f | → | ::=Num +f d -case -wh_rel | Num,-f |
| 1.0000 | +wh_rel c_rel,-wh_rel | → | ::=t +wh_rel c_rel | t,-wh_rel |
| 1.0000 | =Num n | → | ::=Num =Num n | Num |
| 1.0000 | =d +case +dat v,-dat | → | ::=p_to =d +case +dat v | p_to,-dat |
| 1.0000 | =d +case +dat v,-dat,-wh_rel | → | ::=p_to =d +case +dat v | p_to,-dat,-wh_rel |
| 1.0000 | =d +case v | → | ::=p_to =d +case v | p_to |
| 1.0000 | =d +case v,-wh_rel | → | ::=p_to =d +case v | p_to,-wh_rel |
| 0.3922 | =d a | → | ::=A =d a | A |
| 0.4193 | =d a | → | ::=A =d a | ::A |
| 0.1885 | =d a | → | +case =d a,-case | |
| 0.9085 | =d little_v | → | ::=¿v =d little_v | v |
| 0.0915 | =d little_v | → | ::=¿v =d little_v | ::v |
| 1.0000 | =d little_v,-wh_rel | → | ::=¿v =d little_v | v,-wh_rel |
| 1.0000 | =d ven | → | ::=¿v =d ven | v |
| 1.0000 | =d ving | → | ::=¿v =d ving | ::v |
| 1.0000 | A | → | ::deg | ::A |
| 0.9966 | Be,-case | → | ::=a Be | a,-case |
| 0.0034 | Be,-case | → | ::=ving Be | ving,-case |
| 1.0000 | Ce | → | ::=t Ce | t |
| 1.0000 | Have,-case | → | ::=ven Have | ven,-case |
| 0.7016 | Num | → | ::=n Num | ::n |
| 0.2959 | Num | → | ::=n Num | n |
| 0.0025 | Num | → | ::n=¿ Num | ::n |
| 0.9864 | Num,-f | → | ::=n Num | ::n -f |
| 0.0136 | Num,-f | → | ::=n Num | n,-f |
| 1.0000 | Pfor | → | +case Pfor,-case | |
| 1.0000 | Pfor,-wh_rel | → | +case Pfor,-case -wh_rel | |
| 1.0000 | Pin,-wh_rel | → | +case Pin,-case -wh_rel | |
| 1.0000 | Pon,-wh_rel | → | +case Pon,-case -wh_rel | |
| 1.0000 | Pto | → | +case Pto,-case | |
| 1.0000 | Pto -dat | → | +case Pto -dat,-case | |
| 1.0000 | Pto -dat,-wh_rel | → | +case Pto -dat,-case -wh_rel | |
| 1.0000 | Pto,-wh_rel | → | +case Pto,-case -wh_rel | |
| 1.0000 | Pwith,-wh_rel | → | +case Pwith,-case -wh_rel | |
| 1.0000 | a,-case | → | =d a | d -case |
| 1.0000 | c | → | ::=t c | t |
| 1.0000 | c_rel | → | +wh_rel c_rel,-wh_rel | |
| 0.4572 | d -case | → | ::=c_rel d -case | c_rel |
| 0.5428 | d -case | → | ::=Num d -case | Num |
| 1.0000 | d -case -wh_rel | → | +f d -case -wh_rel,-f | |
| 0.3985 | little_v,-case | → | =d little_v | d -case |
| 0.6015 | little_v,-case | → | =d little_v | ::d -case |
| 1.0000 | little_v,-case -wh_rel | → | =d little_v | d -case -wh_rel |
| 0.9925 | little_v,-case,-wh_rel | → | =d little_v,-wh_rel | ::d -case |
| 0.0075 | little_v,-case,-wh_rel | → | =d little_v,-wh_rel | d -case |
| 0.6456 | n | → | ::=Ce n | Ce |
| 0.3544 | n | → | ::A | ::n |
| 1.0000 | n,-f | → | =Num n | Num,-f |
| 1.0000 | p_for | → | ::=¿Pfor p_for | Pfor |
| 1.0000 | p_for,-wh_rel | → | ::=¿Pfor p_for | Pfor,-wh_rel |
| 1.0000 | p_in,-wh_rel | → | ::=¿Pin p_in | Pin,-wh_rel |
| 1.0000 | p_on,-wh_rel | → | ::=¿Pon p_on | Pon,-wh_rel |
| 1.0000 | p_to | → | ::=¿Pto p_to | Pto |
| 1.0000 | p_to,-dat | → | ::=¿Pto p_to | Pto -dat |
| 1.0000 | p_to,-dat,-wh_rel | → | ::=¿Pto p_to | Pto -dat,-wh_rel |
| 1.0000 | p_to,-wh_rel | → | ::=¿Pto p_to | Pto,-wh_rel |
| 1.0000 | p_with,-wh_rel | → | ::=¿Pwith p_with | Pwith,-wh_rel |
| 1.0000 | t | → | +case t,-case | |
| 0.5158 | t,-wh_rel | → | +case t,-case -wh_rel | |
| 0.4842 | t,-wh_rel | → | +case t,-case,-wh_rel | |
| 0.5037 | v | → | ::=Ce v | Ce |
| 0.4120 | v | → | +case v,-case | |
| 0.0823 | v | → | ::=p_for v | p_for |
| 0.0020 | v | → | ::=A v | A |
| 0.2526 | v,-wh_rel | → | ::tmp | v,-wh_rel |
| 0.3806 | v,-wh_rel | → | +case v,-case -wh_rel | |
| 0.1269 | v,-wh_rel | → | +dat v,-dat,-wh_rel | |
| 0.1269 | v,-wh_rel | → | +dat v,-wh_rel,-dat | |
| 0.0908 | v,-wh_rel | → | +case v,-case,-wh_rel | |
| 0.0056 | v,-wh_rel | → | p_in,-wh_rel | v |
| 0.0056 | v,-wh_rel | → | p_with,-wh_rel | v |
| 0.0056 | v,-wh_rel | → | p_on,-wh_rel | v |
| 0.0056 | v,-wh_rel | → | p_for,-wh_rel | v |
| 1.0000 | ven,-case | → | =d ven | ::d -case |
| 1.0000 | ving,-case | → | =d ving | d -case |

Fig. 14. Probabilistic grammar for the Keenan and Hawkins test set.

Weighting the derivations from each of the six stimulus types by these corpus counts (Fig. 13) gives a probabilistic derivation grammar like the one in Fig. 14. By assigning each derivation tree branch a probability, this grammar, which corresponds to the Kaynian promotion analysis, defines a generative probabilistic model for MG derivations, in the same way the grammar in Fig. 1 does.

This model makes it possible to calculate the entropy reduction brought about by a word in a sentence generated by an MG.

Following the method of Fig. 6, this calculation was carried out for the Keenan and Hawkins stimuli[10] by chart-parsing[11] a finite-state representation of an incomplete prefix string, as in Fig. 4.

Then the entropy of the resulting chart, viewed as a PCFG generating the language of situated lexical entries, was calculated by evaluating Equation 1. This yields an uncertainty in bits between every word in each stimulus sentence. Downward changes—the entropy reductions—in these values were then summed for each sentence to arrive at predictions of total difficulty.

## 7. Results

The ERH complexity metric correlates with understandability as measured by Keenan and Hawkins's (1987) repetition accuracy scores; it correctly predicts greater difficulty farther out on the AH (Fig. 15a).

However this correlation only obtains with the grammar expressing the Kaynian promotion analysis, and not on the adjunction analysis (Fig. 15b).[12]

## 8. Discussion

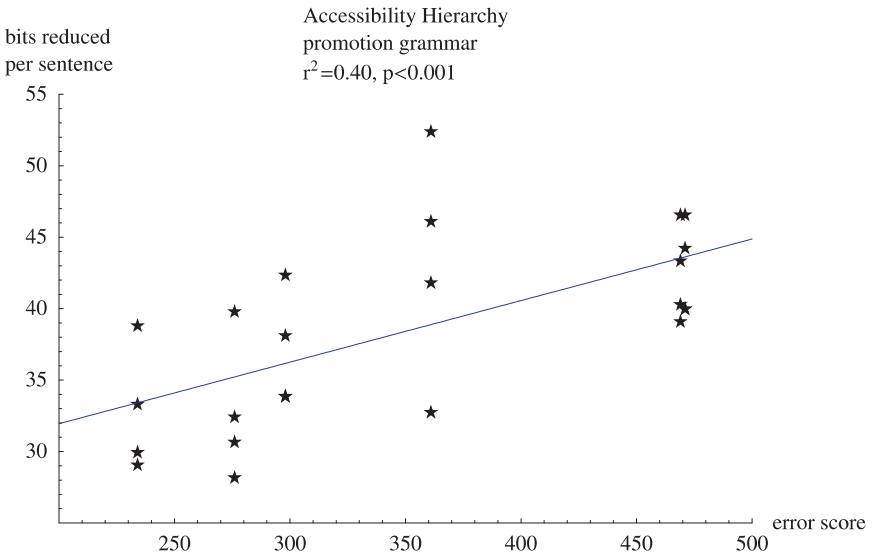
The ERH exemplifies a cognitive explanation for Keenan and Hawkins's (1987) repetition accuracy findings; the true English grammar specifies informationally tougher intermediate states for relative clauses formed on positions lower on the AH. Indeed, the contrasting correlations reported in Section 7 demonstrate that the detailed structure of the grammar plays an important role in the ERH's predictions.

In particular, the promotion and adjunction grammars each define different incremental parser states associated with different degrees of uncertainty about the rest of the sentence. This yields contrasting predictions on some stimuli (subsection 8.2) and convergent predictions on others (subsection 8.1), as discussed next.

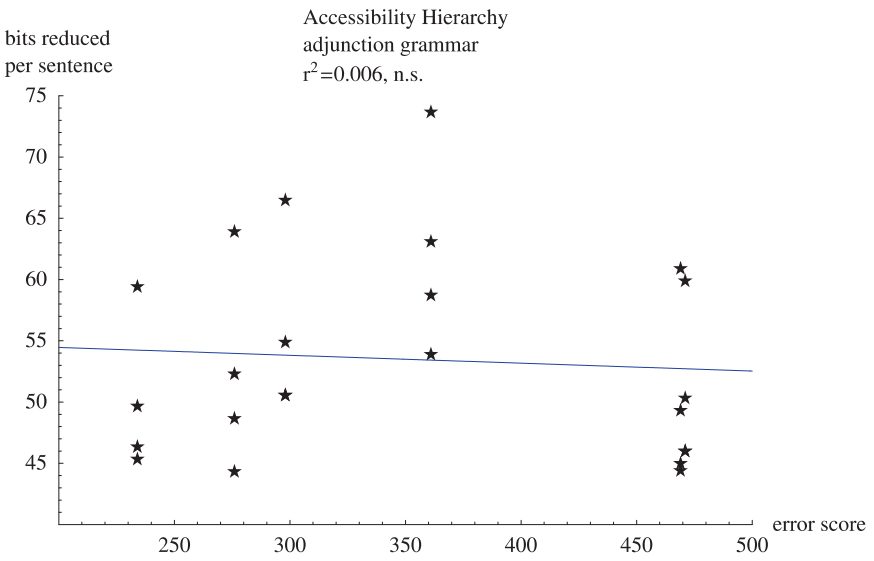
### *8.1. Common predictions*

Even with different relative clause analyses, the two grammars support the same kinds of explanations for certain asymmetries in a processing-based explanation of the AH.

For instance, subject extracted relative clauses are predicted to be easier to understand than indirect object extracted relative clauses on both grammars. This is because, when reading




(a) Per-sentence entropy reductions on the promotion grammar




(b) Per-sentence entropy reductions on the adjunction grammar

Fig. 15. Predictions of two probabilistic MGs in conjunction with the ERH.

stimuli like those in Fig. 16 from left to right, a processor can deduce the relativized grammatical relation in the subject extracted cases immediately. These stimuli never force a transition through a parser state where, for instance, the distinction between direct and indirect object extraction is represented. The derivation nonterminal for the tensed verb phrase resolving this question has an entropy of around 13 bits (on the model in Fig. 14) and never comes into play in the subject extracted cases. Part of this uncertainty stems from the possibility that dative shift could occur with ditransitive verbs.

| *subject* | *indirect object* |
|---|---|
| the boy who told... | the man who Ann gave... |
| the girl who paid... | the boy who Paul sold... |
| the girl who got... | the man who Stephen explained... |
| the man who sold... | the dog which Mary taught... |

Fig. 16. Subject versus Indirect Object prefixes.

The asymmetry that the AH predicts between subject and direct object extraction could in principle be given a similar explanation on the ERH. However, only *give* and *explain* are attested ditransitively in the Keenan and Hawkins (1987) stimuli, so neither grammar actually describes the higher entropy parser states that would make difficulty predictions higher on the other direct object stimuli with *buy* and *write.*

The common treatment of prepositional phrases in both grammars underlies the prediction that relativization from indirect object is easier to understand than relativization from oblique. In both grammars, the verbs *give, sell,* and *teach* take *to*-headed prepositional complements (*sell* can also be merely intransitive). This constrains relativization from indirect object to one kind of phrase-structural launching site. Obliques, on the other hand, are treated as adjuncts whose head (equiprobably *in, with, on,* or *for*) and distance from the verb cannot be predicted.

### *8.2. Contrasting predictions*

The two grammars do not agree on everything. The promotion grammar, for instance, predicts greater difficulty on the sentences in Fig. 17.

they had forgotten that the girl whose friend bought the cake was waiting
I know that the boy whose father sold the dog is very sad
he remembered that the girl whose mother sent the clothes came too late
they had forgotten that the man whose house Patrick bought was so ill
I know that the woman whose car Jenny sold was very angry
he remembered that the girl whose picture Clare showed us was pretty

Fig. 17. Possessives.

These stimuli are just the ones that employ the possessive *whose,* and their contrasting predictions identify a general property of the ERH that plays out differently in the two relative clause analyses.

On Kayne's (1994) analysis, *who* is an independently movable WH-determiner, naturally viewed as a morphological subpart of the combination *who* + *s*. On this analysis, the possessive morpheme *s* has an independent existence, attested in examples such as *whose brother's sister's mother's friend's uncle*. The promotion grammar therefore treats *s* as a recursive category that can combine any pair of number-marked common nouns to yield another noun.[13] This re-

cursive quality leads to more uncertain parser states on the promotion grammar, and contributes to the overall correlation between entropy reduction and repetition error.

By contrast, it is not necessary to morphologically decompose *whose* in Chomsky's (1977) analysis and this lack of recursion keeps parser states comparatively certain at corresponding points in the possessive examples.

The difference in recursivity that shows up in those examples also occurs in another subset of the stimuli, revealing a similar consequence of the treatment of relative clauses. In just the stimuli embedded in the carrier frame *the fact that …* (Fig. 18) the adjunction grammar predicts higher difficulty.

the fact that the girl who paid for the ticket is very poor doesn't matter
the fact that the cat which David showed to the man likes eggs is strange
the fact that the boy who Paul sold the book to hates reading is strange
the fact that the girl who Sue wrote the story with is proud doesn't matter

Fig. 18. Stimuli in carrier frame *the fact that ….*

The adjunction grammar predicts anomalously elevated difficulty on the *fact that* stimuli because it permits recursive modification of any NP, including *the fact* and *the fact that …* with a rule analogous to dP → dP c_relP. This rule means that every expected dP is very uncertain. The promotion grammar, by contrast, does not generalize in this direction, because outermost (vs. stacked) relative clause categories are distinguished by a +f promotion feature. Because only one relative clause is ever stacked in the Keenan and Hawkins (1987) stimulus set, the relevant recursion is not attested, yielding a category of caseless subject dP that is more certain than it is in the adjunction grammar.

In conjunction with the ERH, grammars that predict a wider diversity of more evenly weighted potential continuations at a given point will predict greater processing difficulty at that point. Calculation of the predictions made by the promotion and adjunction grammars highlights the fact that recursive rules—of any type, but especially ones with alternatives that are less sharply biased toward nonrecursive base cases—lead to predictions of heightened difficulty on the ERH.

## 9. Comparison

From a more general perspective, the ERH highlights the interpretation of grammars as definitions of possible continuations. Parser states involving initial substrings with suffixes that are highly uncertain, at a syntactic level, are predicted to be more difficult. This perspective offers an alternative to other processing theories concerned with similar types of data.

### *9.1. Node count*

The most successful previous work modeling performance aspects of the AH is due to J. Hawkins. Hawkins (1994) defined a complexity scale in terms of the cardinality of nodes in

domination, sisterhood, or c-command relations with a relativized structural position. On reasonable grammars, he suggested, this scale corresponds with ease of human processing. This proposal follows Yngve (1960) and Miller and Chomsky (1963) in taking a number of syntactic nodes as a psycholinguistic prediction.

In fact, on the promotion grammar, the total number of derivation tree nodes correlates significantly with AH position. It may be that the ERH emulates Hawkins's proposal by systematically predicting greater difficulty where longer subderivations are possible. In virtue of presupposing a probabilistic grammar, however it breaks with Hawkins by not penalizing larger trees with structure that is very certain.

### *9.2. Integration cost*

Other theories account for relative clause processing asymmetries with costs associated with creating linguistic dependencies across intervening words (Gibson, 2000; Morrill, 2000) or overcoming similarity-based interference from intervening words (Lewis & Vasishth, 2005). Such proposals are compatible with a range of specific disambiguation mechanisms, ranging in bandwidth from single-path reanalysis to any number of ranked parallel parses (Gibson, 1991; Jurafsky, 1996; Narayanan & Jurafsky, 2001; Stevenson & Smolensky, 2006). On the ERH, the cost of updating the set (however large) of viable analyses is influenced by the distribution on unchosen grammatical possibilities, to the extent that they are mutually exclusive, as suggested by Pearlmutter and Mendelsohn (2000).

Although neither Gibson's (2000) approach nor Lewis and Vasishth's (2005) proposals deal explicitly in probabilistic grammars, both could use different construction frequencies or different base activations on grammatical structures to take entropy reduction into account.

### *9.3. Reanalysis cost*

The ERH represents the strong position that any reanalysis happens at the earliest possible point; it is a theory of an eager processor. Specific algorithms that decide when to backtrack (Chater, Crocker, & Pickering, 1998; Stevens & Rumelhart, 1975) are incompatible with it to the extent that they pursue ungrammatical analyses any longer than logically necessary (Tabor, Galantuccia, & Richardson, 2004).

## 10. Conclusion

Further empirical work will be needed to evaluate these theoretical alternatives. The conclusion of this work, however, is that Wilson and Carroll's entropy reduction idea can be extended to infinite languages. This extension can encompass quite expressive grammar formalisms, like MGs, and permits the explicit formulation of a psycholinguistic hypothesis that has been attractive since the 1950s but was until now, unusable. The ERH specified with Equation 2 can be examined quite specifically in combination with a probabilistic grammar. Modeling the results of Keenan and Hawkins (1987) with the ERH leads to a new, more detailed processing explanation for a putative linguistic universal, the AH.

The explanation is that the syntactic structure of, for example, Genitive Object-relativized NPs, is more uncertain during incremental comprehension than the structure of, for instance, Subject-relativized NPs.

This explanation requires some commitment to the linguistic structures involved, as well as their probability. However, because these questions are under consideration in theoretical linguistics as well, it is natural to combine them to yield a unified account.

## Notes

1. The convention that the start symbol is S, that nonterminals are written in capital letters, and that terminals are in lowercase will be adhered to throughout.
2. For example, if there are only two VP rules $\begin{array}{llll} 0.7 & \text{VP} & \rightarrow & \text{V NP} \\ 0.3 & \text{VP} & \rightarrow & \text{V NP NP} \end{array}$ the number of NPs expected when rewriting VP is $0.7 \times 1 + 0.3 \times 2 = 1.3$.
3. In this case, the conditional entropy has the form $H(X|Y = y)$ where $X$ is the random variable whose outcomes that are parses consistent with the observed sentence through word $i - 1$, and $y$ is the actual $i$th word.
4. Keenan and Comrie (1977) defined specific criteria for each grammatical relation. Oblique NPs are typically preceded by prepositions in English, Genitives suffixed with *'s*, and Objects of Comparison (OCOMP) preceded by *than*.
5. Consideration is restricted here to the most well-established part of the AH. Keenan and Hawkins (1987) also considered Object of Comparison, and genitive as well as nongenitive subjects of a passive predicate.
6. The frequency facts alone would not seem to constitute an complete explanation for the AH performance generalization, given that some rare constructions in English are mastered by native speakers. A more adequate explanation would say what it is about processing rare constructions that is so difficult, perhaps by reference to theorized states or operations of the human parser.
7. GenS and GenO stand for Genitive Subject and Genitive Object, respectively.
8. The structural descriptions in Figs. 10 and 11 are generated by particular formal grammars described in Section 5.3. The corresponding derivation trees are shown in Fig. 12. Appendix A concisely presents the grammar formalism and probability model.
9. In Fig. 12, the lower circle corresponds to the movement of "boy," the upper to the movement of the WH-phrase itself, as explained in Section 5.2. Appendix A reviews how MGs realize transformational generalizations in a polynomial-time parsable way. Hale (2003a) discusses the formulation of these analyses as formal grammars in comparatively greater detail.
10. The actual test set was cleaned up in two ways: (a) Because the grammar treats tense and other aspects of the English auxiliary system, test set strings were assumed to be the result of a morphological analyzer that has segmented the words of the actual stimuli. For example, *wrote* is segmented into *write -ed, was* is segmented into *be -ed,* and so on. (b) To eliminate number agreement as a source of derivational uncertainty, four NP in the original Keenan and Hawkins stimuli were changed from plural to singular.

11. To make this process practical, only derivation tree branches observed in correct parses of the test set were ever considered. This artificially restricts the set of continuations to those that can be analyzed using non-zero-weighted branches of the derivation-tree grammar. Because no other MG treebanks exist, statistics on other derivation tree branches are unavailable, and the effect of this approximation cannot be accurately determined.
12. Note that there are four predictions per grammatical function, although in two cases analogous syntactic structure in different stimulus sentences leads to the same prediction, indicated with a single star.
13. McDaniel, McKee, and Bernstein (1998) presented another psycholinguistic application of this analysis.
14. MCFGs also generalize the linear context-free rewriting systems of Weir (1988) by dropping some restrictions on string manipulation functions.

## Acknowledgments

I wish to thank Paul Smolensky, Ed Stabler, and Ted Gibson. The help of Henk Harkema, Roger Levy, and several anonymous reviewers is also gratefully acknowledged.

## References

Alexiadou, A., Law, P., Meinunger, A., & Wilder, C. (Eds.). (2000). *The syntax of relative clauses.* Philadelphia: John Benjamins.

Aoun, J., & Li, Y. (2003). *Essays on the representational and derivational nature of grammar: The diversity of* wh-*constructions*. Cambridge, MA: MIT Press.

Bar-Hillel, Y., Gaifman, C., & Shamir, E. (1960, June). On categorial and phrase-structure grammars. *Bulletin of the Research Council of Israel, 9F,* 1–16.

Bever, T. G. (1970). The cognitive basis for linguistic structures. In J. Hayes (Ed.), *Cognition and the development of language* (pp. 279–362). New York: Wiley.

Bianchi, V. (2000). The raising analysis of relative clauses: A reply to Borsley. *Linguistic Inquiry, 31,* 124–140.

Billot, S., & Lang, B. (1989). The structure of shared forests in ambiguous parsing. In *Proceedings of the 1989 Meeting of the Association for Computational Linguistics.*

Borsley, R. D. (1997). Relative clauses and the theory of phrase structure. *Linguistic Inquiry, 28,* 629–647.

Brame, M. K. (1976). In *Conjectures and refutations in syntax and semantics* (pp. 125–143). Amsterdam: North-Holland.

Bresnan, J. (Ed.). (1982). *The mental representation of grammatical relations*. Cambridge, MA: MIT Press.

Chater, N., Crocker, M. W., & Pickering, M. J. (1998). The rational analysis of inquiry: The case of parsing. In M. Oaksford & N. Chater (Eds.), *Rational models of cognition* (pp. 441–468). New York: Oxford University Press.

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

Chomsky, N. (1956). Three models for the description of language. *IRE Transactions on Information Theory, 2*(3), 113–124.

Chomsky, N. (1977). On Wh-movement. In P. Culicover, T. Wasow, & A. Akmajian (Eds.), *Formal syntax* (pp. 71–132). New York: Academic.

Chomsky, N. (1995). *The minimalist program*. Cambridge, MA: MIT Press.

Frank, R. (2004). Restricting grammatical complexity. *Cognitive Science, 28,* 669–697.

Gibson, E. (1991). *A computational theory of human linguistic processing: Memory limitations and processing breakdown.* Unpublished doctoral dissertation, Carnegie Mellon University, Pittsburgh, PA.

Gibson, E. (1998). Linguistic complexity: Locality of syntactic dependencies. *Cognition, 68,* 1–76.
Gibson, E. (2000). The dependency locality theory: A distance-based theory of linguistic complexity. In Y. Miyashita, A. Marantz, & W. O'Neil (Eds.), *Image, language, brain* (pp. 95–126). Cambridge, MA: MIT Press.
Goldman-Eisler, F. (1958). Speech production and the predictability of words in context. *Quarterly Journal of Experimental Psychology, 10,* 96–106.
Grenander, U. (1967). *Syntax-controlled probabilities* (Tech. Rep.). Providence, RI: Brown University, Division of Applied Mathematics.
Hale, J. (2003a). *Grammar, uncertainty and sentence processing.* Unpublished doctoral dissertation, Johns Hopkins University, Baltimore.
Hale, J. (2003b). The information conveyed by words in sentences. *Journal of Psycholinguistic Research, 32,* 101–123.
Hale, J., & Stabler, E. (2005). Strict deterministic aspects of minimalist grammars. In P. Blache, E. Stabler, J. Busquets, & R. Moot (Eds.), *Logical aspects of computational linguistics* (pp. 162–176). New York: Springer.
Harkema, H. (2001). *Parsing minimalist grammars.* Unpublished doctoral dissertation, University of California at Los Angeles, Los Angeles.
Hawkins, J. A. (1994). *A performance theory of order and constituency* (Vol. 73). New York: Cambridge University Press.
Joshi, A. K. (1985). Tree adjoining grammars: How much context-sensitivity is required to provide reasonable structural descriptions? In D. Dowty, L. Karttunen, & A. M. Zwicky (Eds.), *Natural language parsing: Psychological, computational and theoretical perspectives* (pp. 206–250). New York: Cambridge University Press.
Joshi, A. K., Levy, L. S., & Takahashi, M. (1975). Tree adjunct grammars. *Journal of Computer and System Sciences, 10,* 136–163.
Joshi, A. K., Vijay-Shanker, K., & Weir, D. (1991). The convergence of mildly context-sensitive grammatical formalisms. In *Foundational issues in natural language processing* (pp. 31–81). Cambridge, MA: MIT Press.
Jurafsky, D. (1996). A probabilistic model of lexical and syntactic access and disambiguation. *Cognitive Science, 20,* 137–194.
Kaplan, R. M. (1972). Augmented transition networks as psychological models of sentence comprehension. *Artificial Intelligence, 3,* 77–100.
Kay, M. (1986). Algorithm schemata and data structures in syntactic processing. In B. J. Grosz, K. S. Jones, & B. L. Webber (Eds.), *Readings in natural language processing* (pp. 125–170). San Francisco: Morgan Kaufmann.
Kayne, R. S. (1994). *The antisymmetry of syntax.* Cambridge, MA: MIT Press.
Keenan, E. L., & Comrie, B. (1977). Noun phrase accessibility and universal grammar. *Linguistic Inquiry, 8,* 63–99.
Keenan, E. L., & Hawkins, S. (1987). The psychological validity of the accessibility hierarchy. In E. L. Keenan (Ed.), *Universal grammar: 15 essays* (pp. 60–85). London: Croom Helm.
Kroch, A., & Joshi, A. (1985). *The linguistic relevance of tree adjoining grammar* (Tech. Rep. No. MS-CIS-85-16). Philadelphia: University of Pennsylvania.
Kučera, H., & Francis, W. N. (1967). *Computational analysis of present-day American English.* Providence, RI: Brown University Press.
Lang, B. (1974). Deterministic techniques for efficient non-deterministic parsers. In J. Loeckx (Ed.), *Proceedings of the 2nd Colloquium on Automata, Languages and Programming* (pp. 255–269). Saarbrücken, Germany: Springer.
Lang, B. (1988). Parsing incomplete sentences. In *Proceedings of the 12th International Conference on Computational Linguistics* (pp. 365–371).
Lewis, R., & Vasishth, S. (2005). An activation-based model of sentence processing as skilled memory retrieval. *Cognitive Science, 29,* 375–419.
Marcus, M., Kim, G., Marcinkiewicz, M. A., MacIntyre, R., Bies, A., Ferguson, M., et al. (1994, March). *The Penn Treebank: Annotating predicate argument structure.* Paper presented at the ARPA Human Language Technology Workshop, Plainsboro, NJ.
McDaniel, D., McKee, C., & Bernstein, J. B. (1998). How children's relatives solve a problem for minimalism. *Language, 74,* 308–334.
Michaelis, J. (1998). Derivational minimalism is mildly context-sensitive. In *Proceedings of Logical Aspects of Computational Linguistics '98.* Grenoble.

Michaelis, J. (2001). *On formal properties of minimalist grammars.* Unpublished doctoral dissertation, Potsdam University, Potsdam, Germany.

Miller, G. A., & Chomsky, N. (1963). Finitary models of language users. In R. D. Luce, R. R. Bush, & E. Galanter (Eds.), *Handbook of mathematical psychology* (pp. 419–491). New York: Wiley.

Morrill, G. (2000). Incremental processing and acceptability. *Computational Linguistics, 26*(3) 319–338.

Narayanan, S., & Jurafsky, D. (2001). A Bayesian model predicts human parse preference and reading time in sentence processing. In T. G. Dietterich, S. Becker, & Z. Ghahramani (Eds.), *Advances in Neural Information Processing Systems 14: Proceedings of the 2001 Conference.* Cambridge, MA: MIT Press..

Pearlmutter, N. J., & Mendelsohn, A. A. (2000). *Serial versus parallel sentence comprehension.* Unpublished manuscript, Northeastern University.

Perlmutter, D., & Postal, P. (1974). *Lectures on relational grammar.* Amherst, MA: LSA Linguistic Institute.

Peters, P. S., & Ritchie, R. W. (1973). On the generative power of transformational grammar. *Information Sciences, 6,* 49–83.

Pollard, C., & Sag, I. A. (1994). *Head-driven phrase structure grammar.* Chicago: University of Chicago Press.

Rohde, D. L. (2002). *A connectionist model of sentence comprehension and production.* Unpublished doctoral dissertation, Carnegie Mellon University, Pittsburgh, PA.

Sag, I. A. (1997). English relative clause constructions. *Journal of Linguistics, 33,* 431–484.

Schachter, P. (1973). Focus and relativization. *Language, 49,* 19–46.

Seki, H., Matsumura, T., Fujii, M., & Kasami, T. (1991). On multiple context-free grammars. *Theoretical Computer Science, 88,* 191–229.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal, 27,* 379–423, 623–656.

Shannon, C. E. (1951). Prediction and entropy of printed English. *Bell System Technical Journal, 30,* 50–64.

Shieber, S., & Johnson, M. (1993). Variations on incremental interpretation. *Journal of Psycholinguistic Research, 22,* 287–318.

Stabler, E. (1994). The finite connectivity of linguistic structure. In C. Clifton, L. Frazier, & K. Rayner (Eds.), *Perspectives on sentence processing* (pp. 303–336). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Stabler, E. (1997). Derivational minimalism. In C. Retoré (Ed.), *Logical aspects of computational linguistics* (pp. 68–95). New York: Springer.

Stabler, E. (2004). Varieties of crossing dependencies. *Cognitive Science, 28,* 699–720.

Stabler, E., & Keenan, E. (2003). Structural similarity. *Theoretical Computer Science, 293,* 345–363.

Stevens, A. L., & Rumelhart, D. E. (1975). Errors in reading: An analysis using an augmented transition network model of grammar. In D. A. Norman & D. E. Rumelhart (Eds.), *Explorations in cognition* (pp. 136–155). New York: Freeman.

Stevenson, S., & Smolensky, P. (2006). Optimality in sentence processing. In P. Smolensky & G. Legendre (Eds.), *The harmonic mind: From neural computation to optimality-theoretic grammar: Vol. 2. Linguistic and philosophical implications* (pp. 307–338). Cambridge, MA: MIT Press.

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

Taylor, W. (1953). Cloze procedure: A new tool for measuring readability. *Journalism Quarterly, 30,* 415–433.

Thorndike, E., & Lorge, I. (1968). *The teacher's word book of 30,000 words.* New York: Teachers' College Press.

van Noord, G. (2001). Robust parsing of word graphs. In J-C. Junqua & G. van Noord (Eds.), *Robustness in languages and speech technology* (pp. 205–238). Kluwer.

Wanner, E., & Maratsos, M. (1978). An ATN approach to comprehension. In M. Halle, J. Bresnan, & G. A. Miller (Eds.), *Linguistic theory and psychological reality* (pp. 119–161). Cambridge, MA: MIT Press.

Weir, D. J. (1988). *Characterizing mildly context-sensitive grammar formalisms.* Unpublished doctoral dissertation, University of Pennsylvania, Philadelphia.

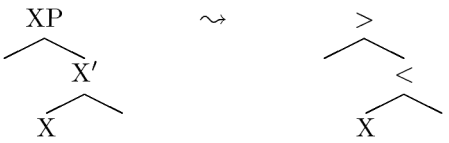
Wilson, K., & Carroll, J. B. (1954). Applications of entropy measures to problems of sequential structure. In C. E. Osgood & T. A. Sebeok (Eds.), *Psycholinguistics: A survey of theory and research* (pp. 103–110). Bloomington: Indiana University Press.

Yngve, V. H. (1960). A model and an hypothesis for language structure. *Proceedings of the American Philosophical Society, 104,* 444–466.

## Appendix A. Minimalist grammars

This appendix outlines the MG formalism as originally defined in Derivational Minimalism (1997). Many equivalent notations for MGs exist, and a more formal presentation can be found in Stabler and Keenan (2003).

MGs generate trees of a certain kind by closing the functions *merge* and *move* on a lexicon of items possessing idiosyncratic features. The trees satisfy a headed $\overline{X}$ theory with only two bar levels: heads and phrases. Notation of these levels is customarily omitted and replaced with individual angle brackets indicating which binary subtree contains the phrase's head.




Each of these trees has a certain string of *features* associated with it. Five feature types are possible.

| | |
|---|---|
| c, t, d, n, v, pred, … | category features |
| =c, =t, =d, =n, =v, =pred, … | selection features |
| +wh, +case, +focus, … | licensor features |
| -wh, -case, -focus, … | licensee features |
| *Lavinia*, *Titus*, *praise*, *-s*, … | phonetic features |

Lexical entries are trees of size zero and are associated with the longest feature strings.

Such trees are called *simple;* all others are *complex*. Phonetic features of empty categories are empty strings ($\epsilon$).

=n d -case *every*
=d =d v *love*
=t +wh c $\epsilon$
…

The structure-building functions operate on these trees and are defined case by case.

*Merge*: If the leftmost feature of the head of $\tau_1$ is =x and the leftmost feature of the head of $\tau_2$ is x, then

$$merge(\tau_1, \tau_2) = \begin{array}{c} < \\ \tau_1' \quad \tau_2' \end{array} \quad \text{if } \tau_1 \text{ is simple, and}$$

$$merge(\tau_1, \tau_2) = \begin{array}{c} > \\ \tau_2' \quad \tau_1' \end{array} \quad \text{if } \tau_1 \text{ is complex,}$$

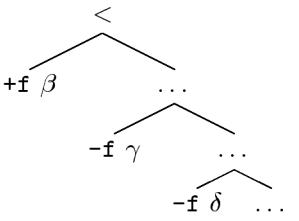
where $\tau_1'$ is like $\tau_1$ except that feature =x is deleted, and $\tau_2'$ is like $\tau_2$ except that feature x is deleted.

*Move*: If the leftmost feature of the head of τ is +y and τ has exactly one maximal subtree $\tau_0$ the leftmost feature of the head of which is -y, then

$move(\tau) =$ 


where $\tau'_0$ is like $\tau_0$ except that feature -y is deleted, and $\tau'$ is like τ except that feature +y is deleted and subtree $\tau_0$ is replaced by a single node without features.

The Shortest Movement Constraint is construed as a mandate not to generate trees with two competing licensee features:




not in domain of *move*

The *closure* of an MG *G* is the set of trees generated by closing the lexicon under the structure-building functions. A *complete* tree is a tree that has only one syntactic feature (e.g., c for complementizer). This is the notion of start category for MGs. Finally, *L*(*G*) the language generated by *G,* is the set of yields of complete trees in the closure of *G*.

### *A.1. Tree-shaped probability model*

MGs are able to generate mildly context-sensitive languages because the *move* rule is nonconcatenative. A fundamental result, obtained independently by Harkema (2001) and Michaelis (2001), is that MGs are equivalent to multiple context-free grammars (MCFGs; Seki et al., 1991). MCFGs generalize standard context-free grammars[14] by allowing the string yields of daughter categories to be manipulated by a function other than simple concatenation. For instance, in a standard CFG, the yield of the parent category is always the string concatenation (⌢) of the yields of the daughters.

| | | | |
|---|---|---|---|
| VP | → | V | NP |
| $s^\frown t$ | ← | $s$ | $t$ |
| S | → | NP | VP |
| $s^\frown t$ | ← | $s$ | $t$ |

This means that if, say, a category deriving a WH-word appears in a verb phrase rule to the right of the verb, then any WH-words derived by that rule will also be to the right of the verb, as shown here.

| | | | |
|---|---|---|---|
| VP | $\rightarrow$ | V | NP |
| *kiss who* | $\leftarrow$ | *kiss* | *who* |
| S | $\rightarrow$ | NP | VP |
| *John kiss who* | $\leftarrow$ | *John* | *kiss who* |

In a nonconcatenative MCFG, the yield of the daughter categories V and NP might be rearranged in some other way.

| | | | |
|---|---|---|---|
| VP | $\rightarrow$ | V | NP |
| $(s, t)$ | $\leftarrow$ | $s$ | $t$ |
| S | $\rightarrow$ | NP | VP |
| $t^\frown r^\frown s$ | $\leftarrow$ | $r$ | $(s, t)$ |

For instance, the string rearranging functions might transport the WH-word to the front of the sentence.

| | | | |
|---|---|---|---|
| VP | $\rightarrow$ | V | NP |
| (*kiss*, *who*) | $\leftarrow$ | kiss | who |
| S | $\rightarrow$ | NP | VP |
| who John kiss | $\leftarrow$ | John | (*kiss*, *who*) |

The power of MCFGs derives from the ability of string handling functions (specified for each rule) to refer to $n$-tuples of categorized strings. This power is restricted by the finitude of these $n$-tuples.

Although too involved to repeat in full detail here, the basis of Harkema and Michaelis's result is the observation that the derivational possibilities for an MG tree structure are entirely determined by three factors: the syntactic features of the head, the tree's status as being simple or complex, and the presence and type of other remaining licensee features in the tree. Coding all this information in a finite way defines the set of relevant categories. These are the classes of trees associated with feature strings that could possibly be generated bottom-up by *merge* and *move*. Because MG lexicons are finite, there are a finite number of licensee feature types per grammar, and hence a finite tuple size required to keep track of them.

With the string manipulation functions fixed by the MG structure-building rules, it becomes possible to give a *derivation tree* for each sentence. Fig. 19 shows a small MG, with the deriva-

| | | | |
|---|---|---|---|
| `=c +nom agrD` | $\epsilon$ | `=i +rel c` | $\epsilon$ |
| `=agrD start` | *the* | `d` | *I* |
| `=d =d i` | *met* | `=n d -rel` | *who* |
| `n -nom` | *boy* | | |

Fig. 19. A minimalist grammar.

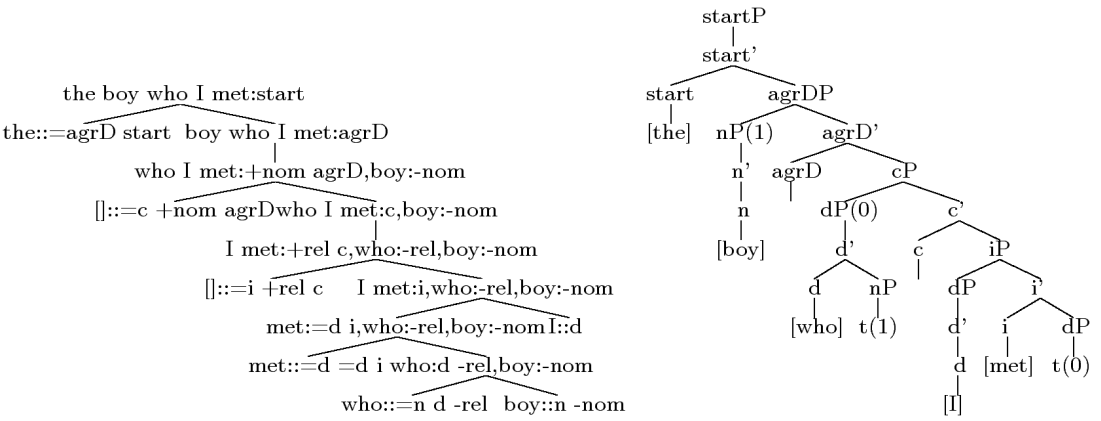



Fig. 20. Derivation tree and structural description of "the boy who I met."

tion tree for "the boy who I met" in Fig. 20 alongside the resulting structural description or derived tree.

As in Tree Adjoining Grammar (Joshi et al., 1975), the nodes of derivation trees like the one on the left in Fig. 20 record instances of tree combination. These derivation trees encode everything there is to know about an MG derivation, and can be parsed in a variety of orders (Harkema, 2001). Most important, if equipped with weights on their branches, they can also be viewed as probabilistic context-free grammars.