The information-processing difficulty of incremental parsing

John Hale
Department of Linguistics and Languages
Michigan State University
East Lansing, MI 48824-1027
jthale@msu.edu

Abstract

When an incremental parser gets the next word, its expectations about upcoming grammatical structures can change. When a word greatly constrains these grammatical expectations, uncertainty is reduced. This elimination of possibilities constitutes information processing work. Formalizing this notion of information processing work yields a complexity metric that predicts human repetition accuracy scores across a systematic class of linguistic phenomena, the Accessibility Hierarchy of relativizable grammatical relations.

1 Introduction

An attractive hypothesis in psycholinguistics, dating back at least to the 1950s, has been that the degree of predictability of words in sentences is somehow related to understandability (Taylor, 1953), production difficulty (Goldman-Eisler, 1958) or, more recently, eye-movements (McDonald and Shillcock, 2003). However, since the 1950s, integrating this hypothesis with realistic models of linguistic structure has remained a challenge.

Lounsbury (1954) appreciated the formal character of the problem. He defined a finite, artificial language, endowed with a rudimentary phonology, morphology and syntax, and showed that a word's informational contribution could be formally defined as the *entropy reduction* brought about by its addition to the end of a sentence fragment. He qualified the significance of his achievement, saying

> An entropy reduction analysis presupposes that the number of possible messages is finite, and that the probabilities 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
>
> (Lounsbury, 1954, 108)

The present paper extends Lounsbury's original idea to infinite 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 context-free phrase structure grammar, and Lang's (1974; 1988) insight that an intermediate parser state is itself a specification of a grammar.

This extension permits the psycholinguistic hypothesis ERH to be examined.

Entropy Reduction Hypothesis (ERH) a person's processing difficulty at a word in a sentence is directly related to the number of bits signaled to the person by that word with respect to a probabilistic grammar the person knows.

In section 2 a method for calculating the entropy reduction of a word in a sentence generated by a probabilistic grammar is presented. Section 3 describes the empirical domain of interest, the Accessibility Hierarchy (Keenan and Comrie, 1977). Section 4 goes on to describe two probabilistic grammars in the class of mildly context-sensitive Minimalist Grammars (Stabler, 1997). One expresses the "promotion analysis" (Kayne, 1994) of relative clauses while the other expresses the more standard "adjunction analysis" (Chomsky, 1977). The predictions of these grammars through the lens of the ERH are considered in sections 5 through 7, where it is shown that predictions derived from the promotion analysis match human repetition accuracy scores better than predictions derived from the adjunction analysis. Section 8 concludes.

2 Entropy Reduction

The idea of the entropy reduction of a word is that uncertainty about grammatical continuations fluctuates as new words come in. The ERH is the proposal that fluctuations in this value be taken as psycholinguistic predictions. This proposal is founded on the possibility of viewing nonterminal symbols

A fuller presentation of this work can be found in Hale (forthcoming).

in probabilistic grammars as random variables. For instance, in the rules given below,

$$\begin{array}{ll} 0.87 & \mathtt{NP} \rightarrow \mathtt{the\ boy} \\ 0.13 & \mathtt{NP} \rightarrow \mathtt{the\ tall\ boy} \end{array}$$

the nonterminal NP can be viewed as a random variable that has two alternative outcomes. Indeed, nonterminals generally in probabilistic context-free phrase structure grammars (PCFGs) can be viewed this way. Since their outcomes are discrete, their entropy H is easily calculated

$$\begin{aligned} H(X) &= -\sum_{x \in X} p(x) \log_2 p(x) \qquad (1) \\ H(\mathtt{NP}) &= -\left[(0.87 \times \log_2 0.87)\right. \\ &\quad \left. +(0.13 \times \log_2 0.13)\right] \\ &\approx 0.56 \text{ bits} \end{aligned}$$

There is just over half a bit of uncertainty about how NP is going to rewrite, because the outcome is so heavily weighted towards the first alternative. In this simple example there is no recursion, so the generated language is finite. To obtain the uncertainty about infinite PCFG languages, a recursive relation due to Grenander (1967) can be used to calculate the entropy of the start symbol S which begins all derivations.

2.1 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 Π and the subset rewriting nonterminal ξ be $\Pi(\xi)$. Denote by p_r the probability of a rule r having daughters $\xi_{j_1}, \xi_{j_2}, \ldots$. Then

$$\begin{aligned} h(\xi_i) &= -\sum_{r \in \Pi(\xi_i)} p_r \log_2 p_r \\ H(\xi_i) &= h(\xi_i) + \sum_{r \in \Pi(\xi_i)} p_r \left[H(\xi_{j_1})\right. \\ &\quad \left. + H(\xi_{j_2}) + \cdots\right] \end{aligned}$$

(Grenander, 1967, 19)

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 to parents.

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

$$H = (I - A)^{-1} \vec{h} \qquad (2)$$

2.2 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 this set, the set of derivations generating a given string $w = w_0 w_1 \ldots w_n$ as a left prefix, is finite, it can be expressed as a list. In the case of a recursive grammar this set is not finite and some other representation is necessary.

Lang and Billot observe (1974; 1988; 1989) that the incremental state of a parser can be described by another, related grammar. They view parsing as the intersection of a grammar with a regular language, of which ordinary strings are but the simplest examples. This perspective readily accommodates incomplete sentences as regular languages whose members all have the same initial n words but continue with all possible words of the terminal vocabulary, for all possible lengths. If $L(G)$ is the language of the grammar G, parsing an initial substring w is the intersection depicted in 3 where the period denotes any terminal symbol of G and the Kleene star indicates any number of repetitions.

$$w(.)^* \cap L(G) \qquad (3)$$

The result of this intersection is a new context-free grammar describing just the derivations whose yield begins with the string w. By generalizing the input from a single string to a regular set of strings, the grammatical continuations can be captured in the new, output grammar. These grammars are easily read off of chart parsers' internal data structures by attaching position indices to nonterminal names, thus distinguishing recognized constituents in different positions.

The uncertainty associated with the the start symbol of this new, resultant grammar is the conditional entropy $H(\mathtt{S}|w_1, w_2, \cdots w_n)$. The entropy reduction of word w_{n+1} then is the downward change in this value as the string w is made one word longer. The proposal of the ERH is that these changes measure the disambiguation work the comprehender has performed by ruling out possible syntactic analyses.

SUBJECT $\supset$ DIR. OBJECT $\supset$ INDIR. OBJECT $\supset$ OBLIQUE $\supset$ GENITIVE $\supset$ OCOMP

Figure 1: The Accessibility Hierarchy of relativizable grammatical relations

3 The Accessibility Hierarchy

This paper examines the processing predictions of the ERH on a systematic class of relative clause types, the Accessibility Hierarchy (AH) shown in figure 1. The AH is an implicational markedness hierarchy of grammatical relations discovered by Keenan and Comrie in (1977). The implication is that if a language has a relative-clause formation rule applicable to grammatical relations at some point x on the AH, then it can also form relative clauses on grammatical relations listed at all points before x.

This hierarchy shows up in a variety of modern syntactic theories that have been influenced by Relational Grammar (Perlmutter and Postal, 1974). In Head-driven Phrase Structure Grammar (Pollard and 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.

Keenan and Comrie speculated that their typological generalization might have a basis in performance factors. This idea was examined in a repetition-accuracy experiment carried out in 1974 but not published until 1987. Subjects in this study repeated back stimulus sentences after a delay while under the additional memory load of a digit-memory task. Stimuli were subject-modifying relative clauses embedded in one of four carrier sentence frames, exemplified in figure 2.

subject extracted they had forgotten that the boy who told the story was so young

direct object extracted the fact that the cat which David showed to the man likes eggs is strange

indirect object extracted I know that the man who Stephen explained the accident to is kind

oblique extracted he remembered that the food which Chris paid the bill for was cheap

genitive subject extracted they had forgotten that the girl whose friend bought the cake was waiting

genitive object extracted the fact that the sailor whose ship Jim took had one leg is important

Figure 2: Relative clauses in each of four carrier sentence types

The results of the human study, given in figure 3, show that repetition accuracy[1] declines across the AH. Keenan and Hawkins (1987) note however that "It remains unexplained just why RCs should be more difficult to comprehend-produce as they are formed on positions lower on the AH."

	SU	DO	IO	OBL	GenS	GenO
repetition accuracy	406	364	342	279	167	171

Figure 3: results from Keenan & Hawkins (1987)

The ERH, if correct, would offer just such an explanation. If a person's difficulty on each word of a sentence is related to derivational information signaled by that word, then the total difficulty reading a sentence ought to be the sum of the difficulty on each word[2].

4 Minimalist Grammars

If correct, the ERH would explain the increasing difficulty across the AH in terms of greater or lesser uncertainty about intermediate parser states. To calculate these predictions, some assumption must be made about what those structures are.

4.1 Two analyses of relativization

Toward this end, two grammars covering the Keenan and Hawkins stimuli were written in the Minimalist Grammars (Stabler, 1997) formalism. These grammars were exactly the same except for their treatment of relative clauses.

One grammar expresses the usual analysis of relative clauses as right-adjoined modifiers (Chomsky, 1977). The other expresses the *promotion* analysis of relative clause. The analysis, which dates back to the 1960s, is revived in Kayne (1994). For reasons having to do with Kayne's general theory of phrase structure, he proposes that, in a sentence like 1, the underlying form of the subject is akin to 2.

[1] Each response was coded for accuracy on a 0-2 scale where 2 means perfect repetition and 1 suggests minor, grammatical errors. A score of 0 was assigned when the response did not include a relative clause of the indicated grammatical function. Cf.Keenan and Hawkins (1987)

[2] Summation naturally extends the word-by-word complexity metric ERH to the sentence level. In word-by-word self-paced reading, evidence for the Accessibility Hierarchy is limited (cf. chapter 5 of Hale (2003)).

(1) the boy who the father explained the answer to was honest

(2) [$_{\mathrm{IP}}$ the father explained the answer to [$_{\mathrm{DP[+wh]}}$ who boy$_{\mathrm{[+f]}}$]]

According to Kayne, at an early stage (2) of syntactic derivation, the determiner phrase (DP) “who boy” occupies what will eventually be the gap position. This DP moves to a specifier position of the enclosing, empty-headed (C0) complementizer phrase (CP), thereby checking a feature +wh as indicated in 3.

(3) [$_{CP}$ [$_{\mathrm{DP}}$ who boy$_{\mathrm{[+f]}}$]$_i$ C0 [$_{\mathrm{IP}}$ the father explained the answer to t$_i$]]

In a second movement, “boy” evacuates from DP, moving to another specifier (perhaps that of the silent agreement morpheme, Agr) as in 4 – checking a different feature, +f.

(4) [$_{\mathrm{AgrP}}$ boy$_j$ Agr [$_{CP}$ [$_{\mathrm{DP}}$ who t$_j$]$_i$ C0 [$_{\mathrm{IP}}$ the father explained the answer to t$_i$]]]

The entire structure becomes a complement of a determiner to yield a larger DP in 5.

(5) [$_{\mathrm{DP}}$ the [$_{\mathrm{AgrP}}$ boy$_j$ Agr [$_{CP}$ [$_{\mathrm{DP}}$ who t$_j$]$_i$ C0 [$_{\mathrm{IP}}$ the father explained the answer to t$_i$]]]]

No adjunction is used in this derivation, and, unconventionally, the leftmost “the” and “boy” do not share an exclusive common constituent. Nor is the wh-word “who” co-indexed with anything. Structural descriptions involving both the Kaynian analysis and the more standard adjunction analysis are shown in figures 4 and 5 respectively[3]. The other linguistic assumptions suggested by these diagrams are discussed in chapter 4 of Hale (2003).

4.2 Formal grammars of relativization

The Minimalist Grammars (MG) formalism (cf. Stabler and Keenan (2003) for a systematic presentation) facilitates the relatively transparent implementation of ideas like movement and feature checking that figure prominently in the two analyses of relativization discussed in the previous subsection. MGs define a set of sentences by closing the structure-building functions *merge* and *move* on a finite set of lexical entries; however, this does not mean that parsing must happen bottom-up. A fundamental result, obtained independently by Harkema (2001) and Michaelis (2001) is that MGs are equivalent to Multiple context-free grammars (Seki et al., 1991). Multiple context-free grammars generalize standard context-free grammars by allowing the string yields of daughter categories to be manipulated by a function other than simple concatenation. As in Tree Adjoining Grammar (Joshi et al., 1975) a record of these manipulations is kept at each node of an MG derivation tree, while a picture of the result is manifested in derived trees such as the ones in figures 4 and 5. The derivation tree on the promotion grammar is shown[4] in figure 6 for the substring “the boy who the father explained the answer to.”

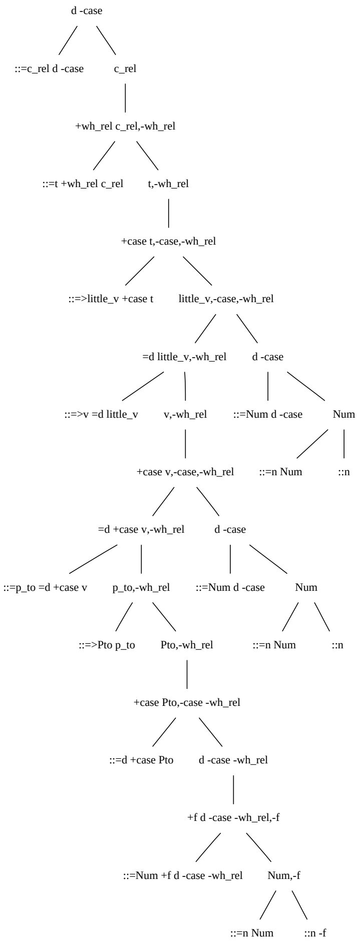

Figure 6: Derivation tree on promotion grammar.

The derivation trees encode everything there is to know about MG derivations, and can be parsed in a variety of orders. Most importantly, if equipped with weights on their branches, they can be generated by probabilistic context-free grammars.

[3] The X-bar structures depicted in figures 4 and 5 are drawn using tools developed by Edward Stabler and colleagues.

[4] These derivation trees are drawn using tools developed by Maxime Amblard.

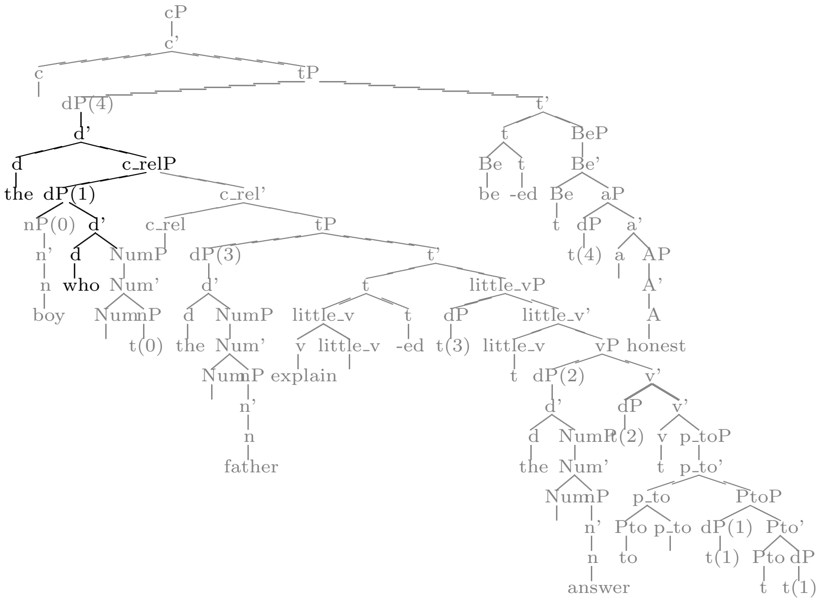
cP
c'
c tP
dP(4) t'
d' t BeP
d c_relP Be t Be'
the dP(1) c_rel' be -ed Be aP
nP(0) d' c_rel tP t dP a'
n' d NumP dP(3) t' t(4) a AP
n who Num' d' t little_vP A'
boy NumP t(0) d NumP little_v t dP little_v' A
the Num' v little_v -ed t(3) little_v vP honest
NumP explain t dP(2) v'
n' d' dP v'
n d NumP t(2) v p_toP
father the Num' t p_to'
NumP p_to PtoP
n' Pto p_to dP(1) Pto'
n to t(1) Pto dP
answer t t(1)

Figure 4: Kaynian promotion analysis

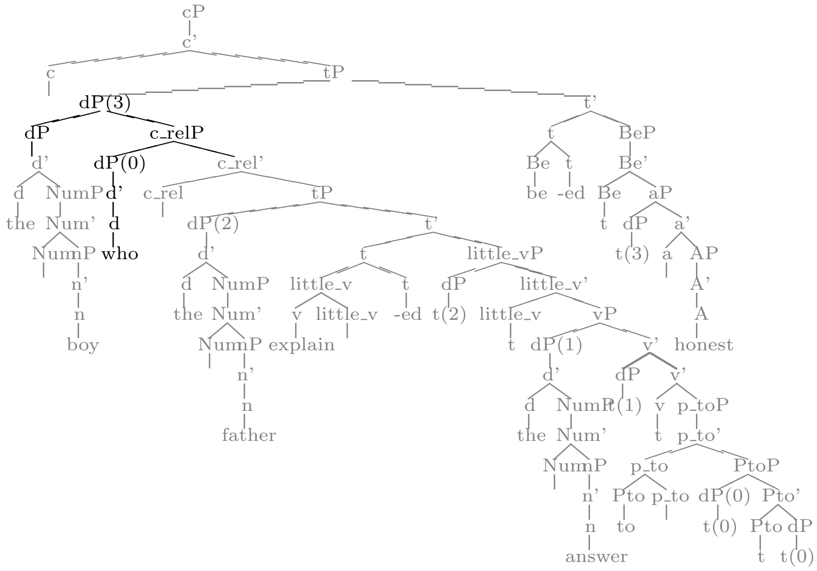
cP
c'
c tP
dP(3) t'
dP c_relP t BeP
d' dP(0) c_rel' Be t Be'
d NumP d' c_rel tP be -ed Be aP
the Num' d dP(2) t' t dP a'
NumP who d' t little_vP t(3) a AP
n' d NumP little_v t dP little_v' A'
n the Num' v little_v -ed t(2) little_v vP A
boy NumP explain t dP(1) v' honest
n' d' dP v'
n d NumP t(1) v p_toP
father the Num' t p_to'
NumP p_to PtoP
n' Pto p_to dP(0) Pto'
n to t(0) Pto dP
answer t t(0)

Figure 5: more standard adjunction analysis

5 Procedure

Derivation trees on both grammars were obtained[5] for each of Keenan and Hawkins' (1987) twenty-four stimulus sentences[6]. Branches of these derivation trees were viewed as PCFG rules with probabilities set according to the usual relative-frequency estimation technique (Chi, 1999). However, because the stimuli were intentionally constructed to have exactly four examples of each structure, these sentences were weighted in accordance with a corpus study (Keenan, 1975) to make their relative frequencies more realistic.

6 Results

The summed entropy reductions exhibit a significant correlation with the repetition accuracy scores collected by Keenan and Hawkins (1987).

The correlation in figure 7(a) obtains only on the grammar expressing the Kaynian promotion analysis, and not on the grammar expressing the standard adjunction analysis (figure 7(b)). Nor do log-probabilities for stimulus sentences on the grammar

[5] Derivations were obtained using a parser described in Appendix A of Hale (2003)

[6] To eliminate number agreement as a source of derivational uncertainty, the results were calculated using a modified stimulus set in which four noun phrases were changed from plural to singular.

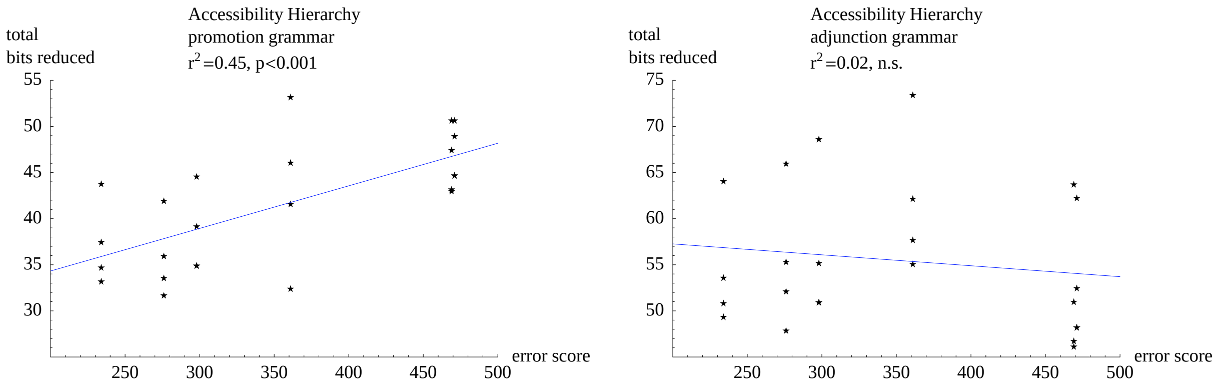

Figure 7: Predictions of two probabilistic Minimalist Grammars through the lens of the ERH

exhibit a significant correlation with repetition accuracy scores.

7 Discussion

From the perspective of the ERH, the difference between the promotion and adjunction grammars resides in the uncertainty of particular states an incremental parser would pass through on the way to a complete analysis.

On the Keenan and Hawkins' (1987) stimuli, these grammars specify incremental parser states that support explanations for some of the observed repetition accuracy asymmetries, abbreviated $<$.

SU $<$ **IO** subject extracted relatives are easier than indirect object extracted relatives, because a left-to-right incremental parser evades, in just subject extracted relatives, the uncertainty associated with questions like

- which internal argument is the gap?
- did dative shift happen?

These questions are defined by alternative derivation-subtrees associated with the verb phrase. For the **DO** stimuli that use potentially ditransitive embedded verbs the same explanation is available, however only two out of four items in the Keenan and Hawkins (1987) set qualify.

IO $<$ **OBL** there is only one type of extraction from indirect object, whereas on these grammars, the head of the oblique phrase ("for" "with" "on" or "in") signals which of four categorically separate kinds of extraction has occurred. These alternatives correspond to four different derivation-nonterminals.

OBL $<$ **GEN** both grammars analyze "whose" as taking a common noun argument, for example "whose ship." But in just the promotion grammar, "whose" is further analyzed as the ordinary "who" morpheme plus a complex possessive phrase headed by "-s" (McDaniel et al., 1998). Because of the recursive character of this possessor category, the structure of "whose's" common noun argument introduces additional uncertainty not present in the indirect object extracted relatives.

Strikingly, the two grammars disagree on six outliers in figure 7(b) where just the adjunction grammar predicts very great difficulty in conjunction with the ERH. These outlier predictions are made on just the sentences that use the nominal carrier frame beginning with "the fact that..." Because the adjunction grammar analyzes relative clauses with an MG rule analogous to the phrase structure rule (4),

$$\mathtt{DP} \longrightarrow \mathtt{DP}\ \mathtt{CP}_{\mathtt{rel}}. \qquad (4)$$

all DPs are available for modification by any number of stacked relative clauses. The nominal frame introduces an additional DP, not present in the other stimuli, that can be modified in this way.

By contrast, the promotion grammar does not include a $+$f promotion feature on any lexical entry for "fact," precluding the possibility of such modification. Moreover, even with such a feature, the promotion grammar assigns different categories to the outermost versus successive relative clause modifiers. 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.

An ERH account that avoids predicting these outliers on the Keenan and Hawkins (1987) stimuli seems to require a grammar where the probability of 2nd and subsequent stacked relative clause modifiers is closer to 0 (its value on the trained promotion grammar) than to 0.31 (its value on the trained adjunction grammar). Beyond these particular stimuli, this modeling motivates a general question about the scale of structural expectations in human sentence processing. Does disconfirmation of a more complicated structural alternative (such as stacked relative clauses) induce greater processing difficulty than disconfirmation of a simpler one? Such empirical issues go beyond the scope of this paper but suggest particular kinds of future work.

8 Conclusion

By extending Lounsbury's (1954) entropy reduction idea to infinite languages, it has become possible to relate predictability and processing difficulty in a way that takes into account linguistic structures defined by one kind of mildly context-sensitive grammar formalism. This relation is the linking hypothesis ERH.

On this linking hypothesis, a grammar expressing the promotion analysis of relative clauses yields whole-sentence predictions more closely approximating human repetition accuracy results than does a grammar expressing the standard adjunction analysis.

If the ERH is true, this result suggests that one grammar carries a kind of greater psychological validity than the other. On the other hand, to the extent that the promotion grammar correctly characterizes human linguistic competence, this confirms the ERH as a linking hypothesis. In any case, the information-processing difficulty of incremental parsing can now be given a more specific definition.

Acknowledgments

The author wishes to thank Paul Smolensky, Ed Stabler and Ted Gibson.

References

Sylvie Billot and Bernard Lang. 1989. The structure of shared forests in ambiguous parsing. In *Proceedings of the 1989 Meeting of the Association for Computational Linguistics*.

Joan Bresnan, editor. 1982. *The Mental Representation of Grammatical Relations*. MIT Press, Cambridge, MA.

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

Noam Chomsky. 1977. On Wh-Movement. In Peter Culicover, Thomas Wasow, and Adrian Akmajian, editors, *Formal Syntax*, pages 71–132. Academic Press, New York.

Frieda Goldman-Eisler. 1958. Speech production and the predictability of words in context. *Quarterly Journal of Experimental Psychology*, 10:96–106.

Ulf Grenander. 1967. Syntax-controlled probabilities. Technical report, Brown University Division of Applied Mathematics, Providence, RI.

John Hale. 2003. *Grammar, uncertainty and sentence processing*. Ph.D. thesis, Johns Hopkins University, Baltimore, Maryland.

Henk Harkema. 2001. *Parsing Minimalist Grammars*. Ph.D. thesis, UCLA.

Aravind K. Joshi, Leon S. Levy, and Masako Takahashi. 1975. Tree adjunct grammars. *Journal of Computer and System Sciences*, 10:136–163.

Richard S. Kayne. 1994. *The Antisymmetry of Syntax*. MIT Press.

Edward L. Keenan and Bernard Comrie. 1977. Noun phrase accessibility and universal grammar. *Linguistic Inquiry*, 8(1):63–99.

Edward L. Keenan and Sarah Hawkins. 1987. The psychological validity of the Accessibility Hierarchy. In Edward L. Keenan, editor, *Universal Grammar: 15 Essays*, pages 60–85, London. Croom Helm.

Edward L. Keenan. 1975. Variation in universal grammar. In R.W. Shuy and R.W. Fasold, editors, *Analyzing Variation in Language*. Georgetown University Press.

Bernard Lang. 1974. Deterministic techniques for efficient non-deterministic parsers. In J. Loeckx, editor, *Proceedings of the 2nd Colloquium on Automata, Languages and Programming*, number 14 in Springer Lecture Notes in Computer Science, pages 255–269, Saarbruücken.

Bernard Lang. 1988. Parsing incomplete sentences. In *Proceedings of the 12th International Conference on Computational Linguistics*, pages 365–371.

Floyd G. Lounsbury. 1954. Transitional probability, linguistic structure and systems of habit-family hierarchies. In C. E. Osgood and T. A. Sebeok, editors, *Psycholinguistics: a survey of theory and research*. Indiana University Press.

Dana McDaniel, Cecile McKee, and Judy B. Bernstein. 1998. How children's relatives solve a problem for minimalism. *Language*, pages 308–334.

Scott A. McDonald and Richard C. Shillcock. 2003. Eye movements reveal the on-line computation of lexical probabilities during reading. *Psychological Science*, 14:648–652.

Jens Michaelis. 2001. *On Formal Properties of Minimalist Grammars*. Ph.D. thesis, Potsdam University.

David Perlmutter and Paul Postal. 1974. *Lectures on Relational Grammar*. LSA Linguistic Institute, UMass Amherst.

Carl Pollard and Ivan A. Sag. 1994. *Head-driven Phrase Structure Grammar*. University of Chicago Press.

Hiroyuki Seki, Takashi Matsumura, Mamoru Fujii, and Tadao Kasami. 1991. On multiple context-free grammars. *Theoretical Computer Science*, 88:191–229.

Edward Stabler and Edward Keenan. 2003. Structural similarity. *Theoretical Computer Science*, 293:345–363.

Edward P. Stabler. 1997. Derivational minimalism. In Christian Retoré, editor, *Logical Aspects of Computational Linguistics*, pages 68–95. Springer.

Wilson Taylor. 1953. Cloze procedure: a new tool for measuring readability. *Journalism Quarterly*, 30:415–433.