A Probabilistic Earley Parser as a Psycholinguistic Model

John Hale
Department of Cognitive Science
The Johns Hopkins University
3400 North Charles Street; Baltimore MD 21218-2685
hale@cogsci.jhu.edu

Abstract

In human sentence processing, cognitive load can be defined many ways. This report considers a definition of cognitive load in terms of the total probability of structural options that have been disconfirmed at some point in a sentence: the surprisal of word w_i given its prefix $w_{0\ldots i-1}$ on a phrase-structural language model. These loads can be efficiently calculated using a probabilistic Earley parser (Stolcke, 1995) which is interpreted as generating predictions about reading time on a word-by-word basis. Under grammatical assumptions supported by corpus-frequency data, the operation of Stolcke's probabilistic Earley parser correctly predicts processing phenomena associated with garden path structural ambiguity and with the subject/object relative asymmetry.

Introduction

What is the relation between a person's knowledge of grammar and that same person's application of that knowledge in perceiving syntactic structure? The answer to be proposed here observes three principles.

Principle 1 *The relation between the parser and grammar is one of strong competence.*

Strong competence holds that the human sentence processing mechanism directly uses rules of grammar in its operation, and that a bare minimum of extragrammatical machinery is necessary. This hypothesis, originally proposed by Chomsky (Chomsky, 1965, page 9) has been pursued by many researchers (Bresnan, 1982) (Stabler, 1991) (Steedman, 1992) (Shieber and Johnson, 1993), and stands in contrast with an approach directed towards the discovery of autonomous principles unique to the processing mechanism.

Principle 2 *Frequency affects performance.*

The explanatory success of neural network and constraint-based lexicalist theories (McClelland and St. John, 1989) (MacDonald et al., 1994) (Tabor et al., 1997) suggests a statistical theory of language performance. The present work adopts a numerical view of competition in grammar that is grounded in probability.

Principle 3 *Sentence processing is eager.*

"Eager" in this sense means the experimental situations to be modeled are ones like self-paced reading in which sentence comprehenders are unrushed and no information is ignored at a point at which it could be used.

The proposal is that a person's difficulty perceiving syntactic structure be modeled by word-to-word surprisal (Attneave, 1959, page 6) which can be directly computed from a probabilistic phrase-structure grammar. The approach taken here uses a parsing algorithm developed by Stolcke. In the course of explaining the algorithm at a very high level I will indicate how the algorithm, interpreted as a psycholinguistic model, observes each principle. After that will come some simulation results, and then a conclusion.

1 Language models

Stolcke's parsing algorithm was initially applied as a component of an automatic speech recognition system. In speech recognition, one is often interested in the probability that some word will follow, given that a sequence of words has been seen. Given some lexicon of all possible words, a *language model* assigns a probability to every string of words from the lexicon. This defines a probabilistic language (Grenander, 1967) (Booth and Thompson, 1973) (Soule, 1974) (Wetherell, 1980).

A language model helps a speech recognizer focus its attention on words that are likely continuations of what it has recognized so far. This is typically done using conditional probabilities of the form

$$P\left(W_n = w_n | W_1 = w_1, \ldots W_{n-1} = w_{n-1}\right)$$

the probability that the nth word will actually be w_n given that the words leading up to the nth have been $w_1, w_2, \ldots w_{n-1}$. Given some finite lexicon, the probability of each possible outcome for W_n can be

estimated using that outcome's relative frequency in a sample.

Traditional language models used for speech are n-gram models, in which $n-1$ words of history serve as the basis for predicting the nth word. Such models do not have any notion of hierarchical syntactic structure, except as might be visible through an n-word window.

Aware that the n-gram obscures many linguistically-significant distinctions (Chomsky, 1956, section 2.3), many speech researchers (Jelinek and Lafferty, 1991) sought to incorporate hierarchical phrase structure into language modeling (see (Stolcke, 1997)) although it was not until the late 1990s that such models were able to significantly improve on 3-grams (Chelba and Jelinek, 1998). Stolcke's probabilistic Earley parser is one way to use hierarchical phrase structure in a language model. The grammar it parses is a probabilistic context-free phrase structure grammar (PCFG), e.g.

$$\begin{array}{ll} 1.0 & \mathrm{S} \rightarrow \mathrm{NP\ VP} \\ 0.5 & \mathrm{NP} \rightarrow \mathrm{Det\ N} \\ 0.5 & \mathrm{NP} \rightarrow \mathrm{NP\ VP} \\ \vdots & \vdots \end{array}$$

see (Charniak, 1993, chapter 5)

Such a grammar defines a probabilistic language in terms of a stochastic process that rewrites strings of grammar symbols according to the probabilities on the rules. Then each sentence in the language of the grammar has a probability equal to the product of the probabilities of all the rules used to generate it. This multiplication embodies the assumption that rule choices are independent. Sentences with more than one derivation accumulate the probability of all derivations that generate them. Through recursion, infinite languages can be specified; an important mathematical question in this context is whether or not such a grammar is *consistent* – whether it assigns some probability to infinite derivations, or whether all derivations are guaranteed to terminate.

Even if a PCFG is consistent, it would appear to have another drawback: it only assigns probabilities to *complete* sentences of its language. This is as inconvenient for speech recognition as it is for modeling reading times.

Stolcke's algorithm solves this problem by computing, at each word of an input string, the prefix probability. This is the sum of the probabilities of all derivations whose yield is compatible with the string seen so far. If the grammar is consistent (the probabilities of all derivations sum to 1.0) then subtracting the prefix probability from 1.0 gives the total probability of all the analyses the parser has disconfirmed. If the human parser is eager, then the "work" done during sentence processing is exactly this disconfirmation.

2 Earley parsing

The computation of prefix probabilities takes advantage of the design of the Earley parser (Earley, 1970) which by itself is not probabilistic. In this section I provide a brief overview of Stolcke's algorithm but the original paper should be consulted for full details (Stolcke, 1995).

Earley parsers work top-down, and propagate predictions confirmed by the input string back up through a set of *states* representing hypotheses the parser is entertaining about the structure of the sentence. The global state of the parser at any one time is completely defined by this collection of states, a *chart*, which defines a tree set. A state is a record that specifies

- the current input string position processed so far
- a grammar rule
- a "dot-position" in the rule representing how much of the rule has already been recognized
- the leftmost edge of the substring this rule generates

An Earley parser has three main functions, *predict*, *scan* and *complete*, each of which can enter new states into the chart. Starting from a dummy start state in which the dot is just to the left of the grammar's start symbol, *predict* adds new states for rules which could expand the start symbol. In these new *predict*ed states, the dot is at the far left-hand side of each rule. After prediction, *scan* checks the input string: if the symbol immediately following the dot matches the current word in the input, then the dot is moved rightward, across the symbol. The parser has "scanned" this word. Finally, *complete* propagates this change throughout the chart. If, as a result of scanning, any states are now present in which the dot is at the end of a rule, then the left hand side of that rule has been recognized, and any other states having a dot immediately in front of the newly-recognized left hand side symbol can now have their dots moved as well. This happens over and over until no new states are generated. Parsing finishes when the dot in the dummy start state is moved across the grammar's start symbol.

Stolcke's innovation, as regards prefix probabilities is to add two additional pieces of information to each state: α, the forward, or prefix probability, and γ the "inside" probability. He notes that

> **path** An (unconstrained) Earley path, or simply path, is a sequence of Earley states linked by prediction, scanning, or completion.

> **constrained** A path is said to be constrained by, or generate a string x if the terminals immediately to the left of the dot in all scanned states, in sequence, form the string x.
>
> ...
>
> The significance of Earley paths is that they are in a one-to-one correspondence with left-most derivations. This will allow us to talk about probabilities of derivations, strings and prefixes in terms of the actions performed by Earley's parser.
>
> (Stolcke, 1995, page 8)

This correspondence between paths of parser operations and derivations enables the computation of the prefix probability – the sum of all derivations compatible with the prefix seen so far. By the correspondence between derivations and Earley paths, one would need only to compute the sum of all paths that are constrained by the observed prefix. But this can be done in the course of parsing by storing the current prefix probability in each state. Then, when a new state is added by some parser operation, the contribution from each antecedent state – each previous state linked by some parser operation – is summed in the new state. Knowing the prefix probability at each state and then summing for all parser operations that result in the same new state efficiently counts all possible derivations.

Predicting a rule corresponds to multiplying by that rule's probability. Scanning does not alter any probabilities. Completion, though, requires knowing γ, the inside probability, which records how probable was the inner structure of some recognized phrasal node. When a state is completed, a bottom-up confirmation is united with a top-down prediction, so the α value of the complete-ee is multiplied by the γ value of the complete-er.

Important technical problems involving left-recursive and unit productions are examined and overcome in (Stolcke, 1995). However, these complications do not add any further machinery to the parsing algorithm per se beyond the grammar rules and the dot-moving conventions: in particular, there are no heuristic parsing principles or intermediate structures that are later destroyed. In this respect the algorithm observes strong competence – principle 1. In virtue of being a probabilistic parser it observes principle 2. Finally, in the sense that *predict* and *complete* each apply exhaustively at each new input word, the algorithm is eager, satisfying principle 3.

3 Parallelism

Psycholinguistic theories vary regarding the amount bandwidth they attribute to the human sentence processing mechanism. Theories of initial parsing preferences (Fodor and Ferreira, 1998) suggest that the human parser is fundamentally serial: a function from a tree and new word to a new tree. These theories explain processing difficulty by appealing to "garden pathing" in which the current analysis is faced with words that cannot be reconciled with the structures built so far. A middle ground is held by bounded-parallelism theories (Narayanan and Jurafsky, 1998) (Roark and Johnson, 1999). In these theories the human parser is modeled as a function from some subset of consistent trees and the new word, to a new tree subset. Garden paths arise in these theories when analyses fall out of the set of trees maintained from word to word, and have to be reanalyzed, as on strictly serial theories. Finally, there is the possibility of total parallelism, in which the entire set of trees compatible with the input is maintained somehow from word to word. On such a theory, garden-pathing cannot be explained by reanalysis.

The probabilistic Earley parser computes all parses of its input, so as a psycholinguistic theory it is a total parallelism theory. The explanation for garden-pathing will turn on the reduction in the probability of the new tree set compared with the previous tree set – reanalysis plays no role. Before illustrating this kind of explanation with a specific example, it will be important to first clarify the nature of the linking hypothesis between the operation of the probabilistic Earley parser and the measured effects of the human parser.

4 Linking hypothesis

The measure of cognitive effort mentioned earlier is defined over prefixes: for some observed prefix, the cognitive effort expended to parse that prefix is proportional to the total probability of all the structural analyses which *cannot* be compatible with the observed prefix. This is consistent with eagerness since, if the parser were to fail to infer the incompatibility of some incompatible analysis, it would be delaying a computation, and hence not be eager. This prefix-based linking hypothesis can be turned into one that generates predictions about word-by-word reading times by comparing the total effort expended before some word to the total effort after: in particular, take the comparison to be a ratio. Making the further assumption that the probabilities on PCFG rules are statements about how difficult it is to disconfirm each rule[1], then the ratio of

[1]This assumption is inevitable given principles 1 and 2. If there were separate *processing* costs distinct from the optimization costs postulated in the grammar, then strong competence is violated. Defining all grammatical structures as equally easy to disconfirm or perceive likewise voids the gradedness of grammaticality of any content.

the α value for the previous word to the α value for the current word measures the combined difficulty of disconfirming all disconfirmable structures at a given word – the definition of cognitive load. Scaling this number by taking its log gives the surprisal, and defines a word-based measure of cognitive effort in terms of the prefix-based one. Of course, if the language model is sensitive to hierarchical structure, then the measure of cognitive effort so defined will be structure-sensitive as well.

5 Plausibility of Probabilistic Context-Free Grammar

The debate over the form grammar takes in the mind is clearly a fundamental one for cognitive science. Much recent psycholinguistic work has generated a wealth of evidence that frequency of exposure to linguistic elements can affect our processing (Mitchell et al., 1995) (MacDonald et al., 1994). However, there is no clear consensus as to the size of the elements over which exposure has clearest effect. Gibson and Pearlmutter identify it as an "outstanding question" whether or not phrase structure statistics are necessary to explain performance effects in sentence comprehension:

> Are phrase-level contingent frequency constraints necessary to explain comprehension performance, or are the remaining types of constraints sufficient. If phrase-level contingent frequency constraints are necessary, can they subsume the effects of other constraints (e.g. locality) ?
>
> (Gibson and Pearlmutter, 1998, page 13)

Equally, formal work in linguistics has demonstrated the inadequacy of context-free grammars as an appropriate model for natural language in the general case (Shieber, 1985). To address this criticism, the same prefix probabilities could be computing using tree-adjoining grammars (Nederhof et al., 1998). With context-free grammars serving as the implicit backdrop for much work in human sentence processing, as well as linguistics[2] simplicity seems as good a guide as any in the selection of a grammar formalism.

6 Garden-pathing

6.1 A celebrated example

Probabilistic context-free grammar (1) will help illustrate the way a phrase-structured language model could account for garden path structural ambiguity. Grammar (1) generates the celebrated garden path sentence "the horse raced past the barn fell" (Bever, 1970). English speakers hearing these words one by one are inclined to take "the horse" as the subject of "raced," expecting the sentence to end at the word "barn." This is the main verb reading in figure 1.

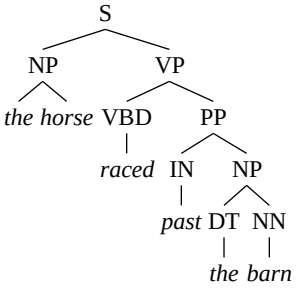


Figure 1: Main verb reading

The human sentence processing mechanism is metaphorically led up the garden path by the main verb reading, when, upon hearing "fell" it is forced to accept the alternative reduced relative reading shown in figure 2.

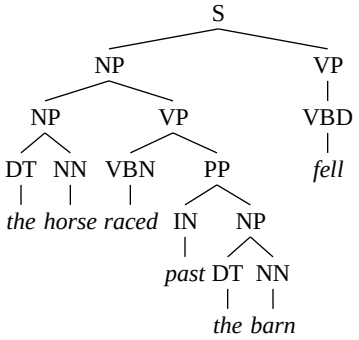


Figure 2: Reduced relative reading

The confusion between the main verb and the reduced relative readings, which is resolved upon hearing "fell" is the empirical phenomenon at issue.

As the parse trees indicate, grammar (1) analyzes reduced relative clauses as a VP adjoined to an NP[3]. In one sample of parsed text[4] such adjunctions are about 7 times less likely than simple NPs made up of a determiner followed by a noun. The probabilities of the other crucial rules are likewise estimated by their relative frequencies in the sample.

[2]Some important work in computational psycholinguistics (Ford, 1989) assumes a Lexical-Functional Grammar where the c-structure rules are essentially context-free and have attached to them "strengths" which one might interpret as probabilities.

[3]See section 1.24 of the Treebank style guide

[4]The sample, starts at sentence 93 of section 16 of the Treebank and goes for 500 sentences (12924 words) For information about the Penn Treebank project see `http://www.cis.upenn.edu/~treebank/`

(1)

1.0	S	→	NP VP .
0.876404494831	NP	→	DT NN
0.123595505169	NP	→	NP VP
1.0	PP	→	IN NP
0.171428571172	VP	→	VBD PP
0.752380952552	VP	→	VBN PP
0.0761904762759	VP	→	VBD
1.0	DT	→	*the*
0.5	NN	→	*horse*
0.5	NN	→	*barn*
0.5	VBD	→	*fell*
0.5	VBD	→	*raced*
1.0	VBN	→	*raced*
1.0	IN	→	*past*

This simple grammar exhibits the essential character of the explanation: garden paths happen at points where the parser can disconfirm alternatives that together comprise a great amount of probability. Note the category ambiguity present with *raced* which can show up as both a past-tense verb (VBD) and a past participle (VBN).

Figure 3 shows the reading time predictions[5] derived via the linking hypothesis that reading time at word n is proportional to the surprisal $\log\left(\frac{\alpha_{n-1}}{\alpha_n}\right)$.

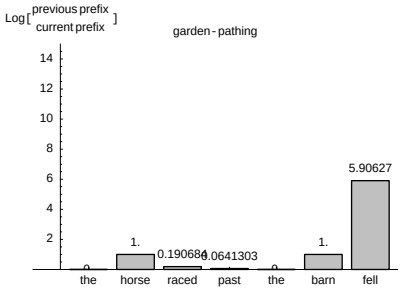


Figure 3: Predictions of probabilistic Earley parser on simple grammar

At "fell," the parser garden-paths: up until that point, both the main-verb and reduced-relative structures are consistent with the input. The prefix probability before "fell" is scanned is more than 10 times greater than after, suggesting that the probability mass of the analyses disconfirmed at that point was indeed great. In fact, all of the probability assigned to the main-verb structure is now lost, and only parses that involve the low-probability NP rule survive – a rule introduced 5 words back.

6.2 A comparison

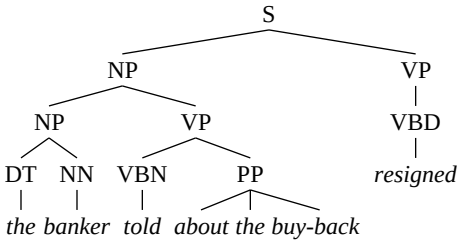
If this garden path effect is truly a result of both the main verb and the reduced relative structures being simultaneously available up until the final verb, then the effect should disappear when words intervene that cancel the reduced relative interpretation early on.

To examine this possibility, consider now a different example sentence, this time from the language of grammar (2).

(2)

0.574927953937	S	→	NP VP
0.425072046063	S	→	VP
1.0	SBAR	→	WHNP S
0.80412371161	NP	→	DT NN
0.082474226966	NP	→	NP SBAR
0.113402061424	NP	→	NP VP
0.11043	VP	→	VBD PP
0.141104	VP	→	VBD NP PP
0.214724	VP	→	AUX VP
0.484663	VP	→	VBN PP
0.0490798	VP	→	VBD
1.0	PP	→	IN NP
1.0	WHNP	→	*who*
1.0	DT	→	*the*
0.33	NN	→	*boss*
0.33	NN	→	*banker*
0.33	NN	→	*buy-back*
0.5	IN	→	*about*
0.5	IN	→	*by*
1.0	AUX	→	*was*
0.74309393	VBD	→	*told*
0.25690607	VBD	→	*resigned*
1.0	VBN	→	*told*

The probabilities in grammar (2) are estimated from the same sample as before. It generates a sentence composed of words actually found in the sample, "the banker told about the buy-back resigned." This sentence exhibits the same reduced relative clause structure as does "the horse raced past the barn fell."



Grammar (2) also generates[6] the subject relative "the banker who was told about the buy-back resigned." Now a comparison of two conditions is possible.

MV and RC *the banker told about the buy-back resigned*

[5] Whether the quantitative values of the predicted reading times can be mapped onto a particular experiment involves taking some position on the oft-observed (Gibson and Schütze, 1999) imperfect relationship between corpus frequency and psychological norms.

[6] This grammar also generates active and simple passive sentences, rating passive sentences as more probable than the actives. This is presumably a fact about the writing style favored by the Wall Street Journal.

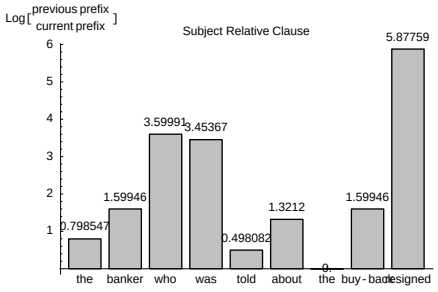


Figure 4: Mean 10.5

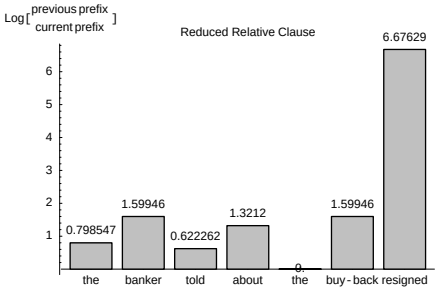


Figure 5: Mean: 16.44

RC only *the banker who was told about the buy-back resigned*

The words *who was* cancel the main verb reading, and should make that condition easier to process. This asymmetry is borne out in graphs 4 and 5. At "resigned" the probabilistic Earley parser predicts less reading time in the subject relative condition than in the reduced relative condition.

This comparison verifies that the same sorts of phenomena treated in reanalysis and bounded parallelism parsing theories fall out as cases of the present, total parallelism theory.

6.3 An entirely empirical grammar

Although they used frequency estimates provided by corpus data, the previous two grammars were partially hand-built. They used a subset of the rules found in the sample of parsed text. A grammar including all rules observed in the entire sample supports the same sort of reasoning. In this grammar, instead of just 2 NP rules there are 532, along with 120 S rules. Many of these generate analyses compatible with prefixes of the reduced relative clause at various points during parsing, so the expectation is that the parser will be disconfirming many more hypotheses at each word than in the simpler example. Figure 6 shows the reading time predictions derived from this much richer grammar.

Because the terminal vocabulary of this richer grammar is so much larger, a comparatively large amount of information is conveyed by the nouns "banker" and "buy-back" leading to high surprisal

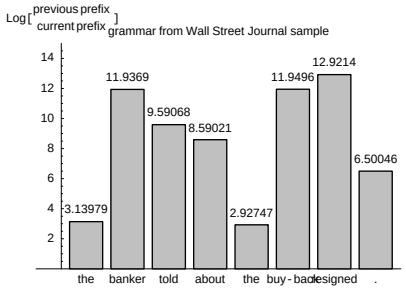


Figure 6: Predictions of Earley parser on richer grammar

values at those words. However, the garden path effect is still observable at "resigned" where the prefix probability ratio is nearly 10 times greater than at either of the nouns. Amid the lexical effects, the probabilistic Earley parser is affected by the same structural ambiguity that affects English speakers.

7 Subject/Object asymmetry

The same kind of explanation supports an account of the subject-object relative asymmetry (cf. references in (Gibson, 1998)) in the processing of unreduced relative clauses. Since the Earley parser is designed to work with context-free grammars, the following example grammar adopts a GPSG-style analysis of relative clauses (Gazdar et al., 1985, page 155). The estimates of the ratios for the two S[+R] rules are obtained by counting the proportion of subject relatives among all relatives in the Treebank's parsed Brown corpus[7].

(3)

0.33	NP	→	SPECNP NBAR
0.33	NP	→	*you*
0.33	NP	→	*me*
1.0	SPECNP	→	DT
0.5	NBAR	→	NBAR S[+R]
0.5	NBAR	→	N
1.0	S	→	NP VP
0.86864638	S[+R]	→	NP[+R] VP
0.13135362	S[+R]	→	NP[+R] S/NP
1.0	S/NP	→	NP VP/NP
1.0	VP/NP	→	V NP/NP
1.0	VP	→	V NP
1.0	V	→	*saw*
1.0	NP[+R]	→	*who*
1.0	DT	→	*the*
1.0	N	→	*man*
1.0	NP/NP	→	ϵ

[7]In particular, relative clauses in the Treebank are analyzed as NP → NP SBAR (rule 1), SBAR → WHNP S (rule 2) where the S contains a trace *T* coindexed with the WHNP. The total number of structures in which both rule 1 and rule 2 apply is 5489. The total number where the first child of S is null is 4768. This estimate puts the total number of object relatives at 721 and the frequency of object relatives at 0.13135362 and the frequency of subject relatives at 0.86864638.

Grammar (3) generates both subject and object relative clauses. S[+R] → NP[+R] VP is the rule that generates subject relatives and S[+R] → NP[+R] S/NP generates object relatives. One might expect there to be a greater processing load for object relatives as soon as enough lexical material is present to determine that the sentence is in fact an object relative[8]. The same probabilistic Earley parser (modified to handle null-productions) explains this asymmetry in the same way as it explains the garden path effect. Its predictions, under the same linking hypothesis as in the previous cases, are depicted in graphs 7 and 8. The mean surprisal for the object relative is about 5.0 whereas the mean surprisal for the subject relative is about 2.1.

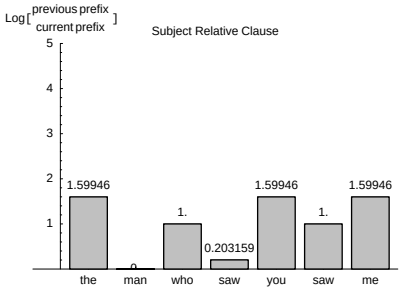


Figure 7: Subject relative clause

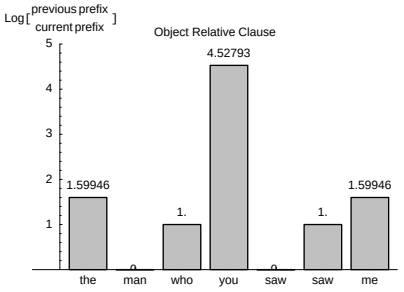


Figure 8: Object relative clause

Conclusion

These examples suggest that a "total-parallelism" parsing theory based on probabilistic grammar can characterize some important processing phenomena. In the domain of structural ambiguity in particular, the explanation is of a different kind than in traditional reanalysis models: the order of processing is not theoretically significant, but the estimate of its magnitude at each point in a sentence is. Results with empirically-derived grammars suggest an affirmative answer to Gibson and Pearlmutter's question: phrase-level contingent frequencies can do the work formerly done by other mechanisms.

Pursuit of methodological principles 1, 2 and 3 has identified a model capable of describing some of the same phenomena that motivate psycholinguistic interest in other theoretical frameworks. Moreover, this recommends probabilistic grammars as an attractive possibility for psycholinguistics by providing clear, testable predictions and the potential for new mathematical insights.

[8]The difference in probability between subject and object rules could be due to the work necessary to set up storage for the filler, effectively recapitulating the HOLD Hypothesis (Wanner and Maratsos, 1978, page 119)

References

Fred Attneave. 1959. *Applications of Information Theory to Psychology: A summary of basic concepts, methods and results*. Holt, Rinehart and Winston.

Thomas G. Bever. 1970. The cognitive basis for linguistic structures. In J.R. Hayes, editor, *Cognition and the Development of Language*, pages 279–362. Wiley, New York.

Taylor L. Booth and Richard A. Thompson. 1973. Applying probability measures to abstract languages. *IEEE Transactions on Computers*, C-22(5).

Joan Bresnan. 1982. Introduction: Grammars as mental representations of language. In Joan Bresnan, editor, *The Mental Representation of Grammatical Relations*, pages xvii,lii. MIT Press, Cambridge, MA.

Eugene Charniak. 1993. *Statistical Language Learning*. MIT Press.

Ciprian Chelba and Frederick Jelinek. 1998. Exploiting syntactic structure for language modelling. In *Proceedings of COLING-ACL '98*, pages 225–231, Montreal.

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

Noam Chomsky. 1965. *Aspects of the Theory of Syntax*. MIT Press, Cambridge MA.

Jay Earley. 1970. An efficient context-free parsing algorithm. *Communications of the Association for Computing Machinery*, 13(2), February.

Janet Dean Fodor and Fernanda Ferreira, editors. 1998. *Reanalysis in sentence processing*, volume 21 of *Studies in Theoretical Psycholingustics*. Kluwer, Dordrecht.

Marilyn Ford. 1989. Parsing complexity and a theory of parsing. In Greg N. Carlson and Michael K. Tanenhaus, editors, *Linguistic Structure in Language Processing*, pages 239–272. Kluwer.

Gerald Gazdar, Ewan Klein, Geoffrey Pullum, and Ivan Sag. 1985. *Generalized Phrase Structure Grammar*. Harvard University Press, Cambridge, MA.

Edward Gibson and Neal J. Pearlmutter. 1998.

Constraints on sentence processing. *Trends in Cognitive Sciences*, 2:262–268.

Edward Gibson and Carson Schütze. 1999. Disambiguation preferences in noun phrase conjunction do not mirror corpus frequency. *Journal of Memory and Language*.

Edward Gibson. 1998. Linguistic complexity: locality of syntactic dependencies. *Cognition*, 68:1–76.

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

Frederick Jelinek and John D. Lafferty. 1991. Computation of the probability of initial substring generation by stochastic context-free grammars. *Computational Linguistics*, 17(3).

Maryellen C. MacDonald, Neal J. Pearlmutter, and Mark S. Seidenberg. 1994. Lexical nature of syntactic ambiguity resolution. *Psychological Review*, 101(4):676–703.

James McClelland and Mark St. John. 1989. Sentence comprehension: A PDP approach. *Language and Cognitive Processes*, 4:287–336.

Don C. Mitchell, Fernando Cuetos, Martin M.B. Corley, and Marc Brysbaert. 1995. Exposure-based models of human parsing: Evidence for the use of coarse-grained (nonlexical) statistical records. *Journal of Psycholinguistic Research*, 24(6):469–488.

Srini Narayanan and Daniel Jurafsky. 1998. Bayesian models of human sentence processing. In *Proceedings of the 19th Annual Conference of the Cognitive Science Society*, University of Wisconsin-Madson.

Mark-Jan Nederhof, Anoop Sarkar, and Giorgio Satta. 1998. Prefix probabilities from stochastic tree adjoining grammars. In *Proceedings of COLING-ACL '98*, pages 953–959, Montreal.

Brian Roark and Mark Johnson. 1999. Broad coverage predictive parsing. Presented at the 12th Annual CUNY Conference on Human Sentence Processing, March.

Stuart Shieber and Mark Johnson. 1993. Variations on incremental interpretation. *Journal of Psycholinguistic Research*, 22(2):287–318.

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

Stephen Soule. 1974. Entropies of probabilistic grammars. *Information and Control*, 25(57–74).

Edward Stabler. 1991. Avoid the pedestrian's paradox. In Robert C. Berwick, Steven P. Abney, and Carol Tenny, editors, *Principle-Based Parsing: computation and psycholinguistics*, Studies in Linguistics and Philosophy, pages 199–237. Kluwer, Dordrecht.

Mark Steedman. 1992. Grammars and processors. Technical Report TR MS-CIS-92-52, University of Pennsylvania CIS Department.

Andreas Stolcke. 1995. An efficient probabilistic context-free parsing algorithm that computes prefix probabilities. *Computational Linguistics*, 21(2).

Andreas Stolcke. 1997. Linguistic knowledge and empirical methods in speech recognition. *AI Magazine*, 18(4):25–31.

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

Eric Wanner and Michael Maratsos. 1978. An ATN approach to comprehension. In Morris Halle, Joan Bresnan, and George A. Miller, editors, *Linguistic Theory and Psychological Reality*, chapter 3, pages 119–161. MIT Press, Cambridge, Massachusetts.

C.S. Wetherell. 1980. Probabilistic languages: A review and some open questions. *Computing Surveys*, 12(4).