

## **Limits on the human sentence generator**

Anthony S. Kroch  
University of Pennsylvania

The problem of language generation is the problem of translating communicative intent into linguistic form. As such, it implicates the entire richness and complexity of human cognition because, to a close approximation, whatever a person can conceive, (s)he can express in language. Computational approaches to generation have, therefore, rightly concentrated on working out, in conjunction with the linguistic disciplines of discourse analysis and pragmatics, the regularities linking communicative intent to linguistic form. In this paper, however, we will address a different problem; namely, the limits on the capacity of the human language generation mechanism to translate preverbal messages into sentences of natural language. Often functional studies of human communicative competence give the impression that people are infinitely subtle and flexible in their ability to use form to signal meaning and intent. But so long as we assume that human behavior is ultimately reduceable to the output of a machine, however complex, it must be the case that there are limits to the competence we are trying to model. If we can find these limits and characterize them theoretically, we will make a contribution to the study of human cognition and will help to circumscribe the problem of language generation by computer. While it is certainly too early in the development of linguistic science to expect a general characterization of the limits on human sentence generation, we hope to show in the brief discussion to follow that some evidence as to their nature is available and that it is reasonable to hope for progress from future empirical work.

### **1. Degrees of difficulty in language processing.**

From the perspective of those of us who would like to develop a computational theory of human language, one of the most puzzling features of people's linguistic behavior is that not all linguistic forms are equally easy for them to process. Moreover, these variations in difficulty cannot be translated into simple differences in the length of time it takes to process different forms, for speakers regularly make errors sensitive to difficulty in processing utterances, both in production and perception. The occurrence of such errors must follow in some way from the finiteness of the computational resources of human beings and from the time constraints under which processing occurs; and accounting for the distribution of such errors poses an obvious challenge to the theory of language use. Of course, some errors, like certain garden path effects in sentence parsing, occur every time the relevant linguistic

environment occurs; and one can quite easily structure a processing algorithm so that it deterministically produces such failures. Indeed, the work on deterministic parsing (Marcus 1980) has shown that one can use the patterning of these errors to choose among competing theories of processing. Other mistakes, however, do not occur every time a given environment appears; but there are stochastic regularities in their distribution. These errors may pose the biggest challenge to computational linguistics and may provide the biggest opportunity for deepening our understanding of the mechanisms underlying sentence processing by human beings.

One example of the sort of error that occurs with measurable frequency in the production of spontaneous utterances in English is the use of resumptive pronouns in relative clauses. Consider, for instance, the following examples which we collected in the course of a recent study of the syntax of relative clauses used in speech (Kroch 1980):

- (1) a. ??I was praying for a lady that she lived near my sister.
- b. ??Let's get the Babar book, the one that B's gonna read it.
- c. ??Frank had an operation on Friday which we just found out about it.
- d. ??You could have a lock on your door that you give your undergraduates a key to it.
- e. ??I have the ones that she felt she couldn't do anything with them.

In each of these cases the relative clause sounds odd because a pronoun appears in a position where English requires a gap or 'empty category'. Of course, there are certain environments, like those in (2) below, in which resumptive pronouns are more acceptable because the grammar of English does not permit a gap in the position of the pronoun:

- (2) a. People are coming out with symptoms that the doctors don't know what they are.
- b. I'll bring a bottle of that stuff that you and I are the only ones that like it.

Whether these latter cases should be allowed by the grammar or classed as ungrammatical with the examples in (1) is not clear; in any event our concern is with the clearly unacceptable cases, which, somewhat surprisingly, occur with measurable frequency in speech. We found in our study that about 1% of relative clauses in spontaneous natural discourse contained resumptive pronouns. Why these pronouns should occur is, of course, hard to determine; but it is possible to construct a plausible mechanism for generating them that is worthy of further investigation. Thus, consider a sentence generating algorithm (e.g., that in MacDonald 1980) under which relative clause gaps are created by zeroing a message element under identity with the head of the relative clause being produced (Kroch 1981). If that zeroing is subject to random failure, then the identical element will occasionally appear in the output string. It will usually be a pronoun, because the mechanism responsible for pronominalization will recognize the repeated

message element as a second reference to the discourse entity that the head of the relative clause points to. Of course, since pronominalization is not obligatory, we might expect to find cases in which the resumptive element is a full NP rather than a pronoun; and such examples, in fact, also occur. The relative clause in (3) below is one such case:

- (3) ??In the middle of the country is a high density area that most of the people come from that area.

It is interesting to note that under a production algorithm that employs zeroing of an identical element rather than movement of the relativized NP to the beginning of the clause, there is no need to fully plan the syntax of a relative clause before beginning to send a partially formulated clause to the output device that turns the syntactic structure into speech. In particular, the syntactic position of the element to be zeroed need not be calculated because the element will be zeroed as it is encountered. If the conditions for zeroing of the identical element are not met, generation can simply proceed; and a resumptive pronoun or NP will appear. Just this seems to happen in natural spoken English in cases where leaving a gap would violate the conditions on empty categories. If we say that zeroing of the identical element is explicitly blocked in these cases, then essentially the same mechanism that explains the sporadic occurrence of clauses like those in (1) above will account for why spoken English commonly exhibits clauses like those in (2). In this respect, English appears to differ from languages like German or the Slavic languages, which must have a somewhat different algorithm for producing relative clauses. Because German and Slavic relative pronouns are marked for case, the speakers of those languages cannot begin to send a relative clause to the output device until the syntactic position of the gap is fixed since it is this syntactic position which determines the case of the relative pronoun. Under these circumstances, it should be much easier for speakers to avoid producing relative clauses with gaps in the wrong position in these languages than in English; and hence resumptive pronouns should be less common in speech in these languages. While no firm evidence on the frequency of resumptive pronouns in German or similar languages is available, experienced observers seem to agree that the sorts of resumptive pronouns that are heard in English do not occur with any noticeable regularity in these other languages. In contrast, western European languages that share with English the property of having minimal case marking on their relativizers (e.g., the Romance languages and the Scandinavian languages) do exhibit use of resumptive pronouns in speech.

If the random occurrence of resumptive pronouns in spoken English relatives is due to the structure of the production algorithm and to a characteristic way in which it can fail, then it should be of interest to find out under what conditions failure is more or less likely to occur. Information on this point can be obtained by statistical comparison of randomly sampled corpora of relative clauses with and without resumptive pronouns. When we carried out such a study on a corpus of 500 relative clauses containing resumptive pronouns collected from naturally occurring discourse and 5000 clauses without resumptive pronouns collected from tape recorded sociolinguistic interviews, we found that one of the most significant factors influencing the likelihood of appearance of a resumptive pronoun was the degree of embedding of the gap position. Thus, if the gap position was in a subordinate clause within the relative clause (as in (1e) above) it was more likely to be filled with a resumptive pronoun than if it was in the highest clause of the relative (as in (1a)-(1d)). Even more strikingly, in simple sentence relative clauses, the likelihood of occurrence of a resumptive pronoun increased with each increase in the number of phrasal nodes on the path between the head of the clause and the gap position. In other words, subject position resumptive pronouns like (1a) were less likely than direct object resumptive pronouns like (1b), and these were less likely than resumptive pronoun objects of verb phrase prepositions (as in (1c)). The most likely position for the occurrence of resumptive pronoun was the position of complement to a direct object NP (as in (1d)), this being the position with the longest path between the gap position and the antecedent. The following table gives probability weights for each of degree of embedding calculated for the sample we analyzed using the VARBRUL 2S program for multivariate logit analysis (Rousseau and Sankoff 1978):

	Likelihood of occurrence of resumptive pronoun
Clausal embedding of gap position:	
in highest clause of relative	.18
in infinitival complement	.62
in tensed complement	.73
Gap position within the clause:	
subject	.30
direct object	.50
object of verb phrase PP	.68
complement to object noun	.93

Table 1: Effect of degree of embedding of gap position on likelihood of occurrence of a resumptive pronoun. (Weights lie between 0.0 and 1.0, with higher weights indicating increased likelihood of occurrence of a resumptive pronoun.)

The challenge posed by these results, which from a statistical point of view are quite robust, is to construct a model of the sentence generation process in which stochastic effects of complexity have a natural place.

## 2. Limits to planning.

One of the limits on language processing that follows from the time and resource constraints under which it operates is that the planning of sentences in generation cannot take account of every conceivably relevant fact about the discourse situation. At some point decisions must be made which cut short the planning process. If this were not so, we would expect, among other things, the forms of sentences used in discourse to be determined by arbitrarily complex predicates on prior discourse context, which certainly seems not to be true. Saying that there must be limits to planning, however, is a great deal easier than showing what these limits are because of the great expressiveness and flexibility of human discourse competence. Nevertheless, it is possible to find evidence that certain mechanical effects - that is, effects not related to meaning or appropriateness - influence syntactic choices by speakers. For instance, in an interesting statistical study of the use of the agentless passive in spontaneous discourse, Weiner and Labov (1983) found that an important factor influencing speaker's choices between active sentences with generalized subjects like (4) and agentless passives like (5) was whether they had used a passive sentence in the preceding 5 clauses:

- (4) They broke into the liquor cabinet.
- (5) The liquor cabinet got broken into.

Weiner and Labov suggested that this result was a 'mechanical' syntactic effect which showed the limits of considerations of discourse function in determining syntactic usage in spontaneous speech. However, while these results and interpretation were intriguing, it was clear to students of discourse function that alternative explanations of the so-called mechanical effect, which might be considered a 'priming' effect in syntax akin to the well-known lexical priming effect, were possible. In particular, it was possible that the effect was an artifact of discourse functional effects not properly controlled for in the study. In order to test the validity of the Weiner and Labov finding, Dominique Estival and I planned a study in which the relevant discourse effects due to topicality of logical subject and object, repetition, aspect, and other factors were explicitly controlled for. We also decided to test whether the priming effect, if it did exist, was sensitive to the difference between verbal and adjectival passives since the two forms of passive, illustrated below in (6) and (7), had been argued convincingly not to be the same syntactic

construction (Wasow 1977):

- (6) John was fired by his boss.
- (7) John was interested in music.

The results of a statistical study of a corpus of more than 600 passive sentences and a roughly equal sized random subsample of active sentences (see Estival 1982, 1985) showed that the priming effect was orthogonal to the discourse function effects controlled for and that it cleanly differentiated verbal from adjectival passives. Note that the probabilistic weights in table 2 below are highest along the main diagonal, which shows that verbal passives are priming verbal passives and adjectival passives are priming adjectival passives, but that verbal and adjectival passives are not priming one another:

	Active	Likelihood of Verbal passive	adjectival passive
Clause type found in preceding 5 clauses:			
active only	.44	.26	.30
verbal passive	.22	.56	.22
adjectival passive	.32	.21	.47

Table 2: Effect of the occurrence of preceding verbal and adjectival passives on the likelihood of a passive.

From these results we concluded that the mere fact that a speaker uses a construction seems to increase the likelihood that (s)he will use it again and hence that the use of syntactic constructions is conditioned, not just by discourse appropriateness but also by their 'accessibility' or 'level of psychological activation' for the sentence production mechanism (see Bock 1982 for further discussion). The fact that the priming effect differentiates verbal from adjectival passives suggests, moreover, that the identity criteria for priming reflect quite abstract linguistic properties.

Another recent study of ours which points out both the flexibility of human sentence planning and the limits on that flexibility is a quantitative study of 700 transitive particle verbs. We investigated the factors which influenced the position of the direct object in these sentences, which, as the examples in (8) and (9) below illustrate, may be either before or after the particle:

- (8) The boy put the dog down.
- (9) The boy called up his friend.

Two of the factors which heavily influenced the relative order of direct object and particle, already known to us from a previous study (Kroch and Small 1978), were the length of the direct object NP and the semantic contribution of the particle to the sentence meaning. The first effect was that longer object NP's were more likely to appear after the particle than shorter ones. The semantic effect was that particles which made an independent contribution to the meaning of the sentence, as, for example, in (8) above, were more likely to occur in post-object position (where equivalent prepositional phrases occur) than purely idiomatic particles like the one in (9). These effects were exactly as expected, given the results of our previous study. We were curious to know, however, whether the two effects were independent of one another; in particular, we wanted to know what happened in cases where the length effect and the semantic effect cut against one another, the relevant case being that of sentences with long direct objects and non-idiomatic particles. The following table shows gives the cross-tabulation of these two factors:

Semantic type of particle	Length of direct object NP		
	1-2 words	3-4 words	5 or more
<b>Idiomatic:</b>			
particle before object NP	114	45	45
total cases	240	62	47
percent particle before NP	48	73	96
<b>Compositional:</b>			
particle before object NP	79	18	9
total cases	263	65	13
percent particle before NP	30	28	69

Table 3: Crosstabulation of particle type by direct object length, showing effects on order of particle and NP and on number of cases of each type.

It is obvious from inspection that the cell which crosses non-idiomatic particles with direct objects five or more words long is much smaller than expected. The first and second columns are roughly the same size in each row; but in the third column, which represents the figures for long NP's, the cell with the figures for idiomatic particles is three times as large as the cell representing the non-idiomatic ones. It appears, therefore, that speakers are avoiding the use of the verb-particle construction in the case where the conditioning factors favor opposite orderings of particle and object NP. Of more interest for our present discussion, however, is that those sentences which do occur in this cell show exactly the intermediate frequency of object last order that we would expect if the two effects were independent of one another. This result is consistent with a model of the production of these sentences in which the decision on how to order the particle and direct object was unaffected by the decision as to whether to use the verb-particle construction at all. It is as though the production mechanism were organized into a simple decision tree in

which the decision as to whether to use a particle verb is made first, apparently on the basis of information about the semantic relationship between verb and particle and about the 'heaviness' (perhaps the amount of descriptive content) of the object NP, and then the particle object ordering decision is made independently. To the extent that such a simple organization of decisions for sentence production, without complex interactions among levels, can be justified by further work, it will be possible to construct a more constrained model of the generation process; and we will have a better idea of the structural characteristics of the system within which discourse functional considerations have their effects.

### References

- Bock, J. K. (1982) "Toward a Cognitive Psychology of Syntax: Information Processing Contributions to Sentence Formulation." Psychological Review 89:1-47.
- Estival, D. (1982) "Analyzing the Passive: How Many Types are There?" in Penn Review of Linguistics, no. 7.
- Estival, D. (1985) "Syntactic Priming of the Passive." in T. Givón, ed. Quantified Studies in Discourse, special issue of Text, 5:7-24.
- Kroch, A. (1980) "Resumptive Pronouns in English Relative Clauses." paper presented at Linguistic Society of America annual meeting.
- Kroch, A. (1981) "On the Role of Resumptive Pronouns in Amnestying Island Constraint Violations." in The Proceedings of the 17th Annual Meeting of the Chicago Linguistics Society.
- Kroch, A. and C. Small (1978) "Grammatical Ideology and its Effect on Speech." in D. Sankoff, ed. Linguistic Variation: Models and Methods. New York: Academic Press.
- MacDonald, D. (1980) Natural Language Production as a Process of Decision-Making under Constraint. MIT Dissertation.
- Marcus, M. (1980) A Theory of Syntactic Recognition for Natural Language. Cambridge: MIT Press.
- Rousseau, P. and D. Sankoff. (1978) "Advances in Variable Rule Methodology." in D. Sankoff, ed. Linguistic Variation: Models and Methods. New York: Academic Press.
- Wasow, T. (1977) "Transformations and the Lexicon." in P. Culicover, T. Wasow, and A. Akmajian, eds. Formal Syntax. New York: Academic Press.
- Weiner, E. J. and W. Labov (1983) "Constraints on the Agentless Passive." Journal of Linguistics, 19: 29-58.