COPYING IN NATURAL LANGUAGES, CONTEXT-FREENESS, AND QUEUE GRAMMARS

Alexis Manaster-Ramer
University of Michigan
2236 Fuller Road #108
Ann Arbor, MI 48105

ABSTRACT

The documentation of (unbounded-length) copying and cross-serial constructions in a few languages in the recent literature is usually taken to mean that natural languages are slightly context-sensitive. However, this ignores those copying constructions which, while productive, cannot be easily shown to apply to infinite sublanguages. To allow such finite copying constructions to be taken into account in formal modeling, it is necessary to recognize that natural languages cannot be realistically represented by formal languages of the usual sort. Rather, they must be modeled as families of formal languages or as formal languages with indefinite vocabularies. Once this is done, we see copying as a truly pervasive and fundamental process in human language. Furthermore, the absence of mirror-image constructions in human languages means that it is not enough to extend Context-free Grammars in the direction of context-sensitivity. Instead, a class of grammars must be found which handles (context-sensitive) copying but not (context-free) mirror images. This suggests that human linguistic processes use queues rather than stacks, making imperative the development of a hierarchy of Queue Grammars as a counterweight to the Chomsky Grammars. A simple class of Context-free Queue Grammars is introduced and discussed.

Introduction

The claim that at least some human languages cannot be described by a Context-free Grammar no matter how large or complex has had an interesting career. In the late 1960's it might have seemed, given the arguments of Bar-Hillel and Shamir (1960) about *respectively* coordinations in English, Postal (1964) about reduplication-cum-incorporation of object noun stems in Mohawk, and Chomsky (1963) about English comparative deletion, that this claim was firmly established.

Potentially serious—and at any rate embarrassing—problems with both the formal and the linguistic aspects of these arguments kept popping up, however (Daly, 1974; Levelt, 1974), and the partial fixes provided by Brandt Corstius (as reported in Levelt, 1974) for the *respectively* arguments and by Langendoen (1977) for that as well as the Mohawk argument did not deter Pullum and Gazdar (1982) from claiming that "it seems reasonable to assume that the natural languages are a proper subset of the infinite-cardinality CFL's, until such time as they are validly shown not to be". Two new arguments, Higginbotham's (1984) one involving *such that* relativization and Postal and Langendoen's (1984) one about sluicing were dismissed on grounds of descriptive inadequacy by Pullum (1984a), who, however, suggested that the Langendoen and Postal (1984) argument about the doubling relativization construction may be correct (all these arguments deal with English).

Pullum (1984b) likewise heaped scorn on my argument that English **reshm**uplicative constructions show non-CFness, but he accepted (1984a; 1984b) Culy's (1985) argument about noun reduplication in Bambara and Shieber's (1985) one about Swiss German cross-serial constructions of causative and perception verbs and their objects. Gazdar and Pullum (1985) also cite these two, as well as an argument by Carlson (1983) about verb phrase reduplication in Engenni. They also refer to my discovery of the **X** *or no* **X** ... construction in English and mention that "Alexis Manaster-Ramer ... in unpublished lectures finds reduplication constructions that appear to have no length bound in Polish, Turkish, and a number of other languages". While they do not refer to my 1983 **reshm**uplication argument, which they presumably still reject, the Turkish construction they allude to was cited in my 1983 paper and is similar to the English **reshm**uplication in form as well as function (see below).

In any case, the acceptance of even one case of non-CFness in one natural language by the only active advocates of the CF position would seem to suffice to remove the issue from the agenda. Any additional arguments, such as Kac (to appear), Kac, Manaster-Ramer, and Rounds (to appear), and Manaster-Ramer (to appear a; to appear b) may appear to be no more than flogging of dead horses. However, as I argued in Manaster-Ramer (1983) and as recent work (Manaster-Ramer, to appear a; Rounds, Manaster-Ramer, and Friedman, to appear) shows ever more clearly, this conception of the issue (viz., Is there one natural languages that is weakly noncontext-free?) makes very little difference and not much sense.

First of all, if non-CFness is so hard to find, then it is presumably linguistically marginal. Second, weak generative arguments cannot be made to work for natural languages, because of their high degree of structural ambiguity and the great difficulty in excluding every conceivable interpretation on which an apparently ungrammatical string might turn out—on reflection—to be in the language. Third, weak generative capacity is in any case not a very interesting property of a formal grammar, especially from a linguistic point of view, since linguistic models are judged by other criteria (e.g., natural languages might well be regular without this making CFGs any the more attractive as models for them). Fourth, results about the place of natural languages in the Chomsky Hierarchy seem to be should be considered in light of the fact that there is no reason to take the Chomsky Hierarchy as the appropriate formal space in which to look for them. Fifth, models of natural languages that are actually in use in theoretical, computational, and descriptive linguistics are —and always have been—only remotely related to the Chomsky Grammars, which means that results about the latter may be of little relevance to linguistic models.

As I argued in 1983, we should go beyond piecemeal debunking of invalid arguments against CFGs and by the same token it seems to me that we must go beyond piecemeal restatements of such arguments. Rather, we should focus on general issues and ones that have implications for the modeling of human languages. One such issue is, it seems to me, the kind of context-sensitivity found in natural languages. It appears that the counterexamples to context-freeness are all rather similar. Specifically, they all seem to involve some kind of cross-serial dependency, i.e., a dependency between the nth elements of two or more substrings. This—unlike the statement that natural languages are noncontext-free—might mean something if we knew what kinds of models were appropriate for cross-serial dependencies. Given that not every kind of context-sensitive construction is found in human languages, it should be clear that there is nothing to be gained by invoking the dubious slogan of context-sensitivity.

Another relevant question is the centrality or peripherality of these constructions in natural languages. The relevant literature makes it appear that they are somewhat marginal at best. This would explain the tortured history of the attempts to show that they exist at all. However, this appears to be wrong, at least when we consider copying constructions. The requirement of full or near identity of two or more subparts of a sentence (or a discourse) is a very widespread phenomenon. In this paper, I will focus on the copying constructions precisely because they are so common in human languages.

In addition to such questions, which appear to focus on the linguistic side of things, there are also the more mathematical and conceptual problems involved in the whole enterprise of modeling human languages in formal terms. My own belief is that both kinds of issues must be solved in tandem, since we cannot know what kind of formal models we want until we know what we are going to model, and we cannot know what human languages are or are not like until we know how to represent them and what to compare them to. This paper is intended as a contribution to this kind of work.

Copying Dependencies

The examples of copying (and other) constructions which have figured in the great context-freeness debate have all involved attempts to show that a whole (natural) language is noncontext free. Now, while it is often easy to find a noncontext-free subset of such a language, it is not always possible to isolate that subset formally from the rest of the language in such a way as to show that the language as a whole is noncontext-free. There is so much ambiguity in natural languages that it is strictly speaking impossible to isolate *any* construction at the level of strings, thus invalidating *all* arguments against CFGs or even Regular Grammars that refer to weak generative capacity. However, the arguments can be reconstructed by making use of the notion of classificatory capacity of formal grammars, introduced in Manaster-Ramer (to appear a) and Manaster-Ramer and Rounds (to appear). The classificatory capacity is the set of languages generated by the various subgrammars of a grammar, and if we are willing to assume that linguists can tell which sentences in a language exemplify the same or different syntactic patterns, then we can usually simply demonstrate that, e.g., no CFG can have a subgrammar generating all and only the sentences of some particular construction if that construction involves reduplication. This will show the inadequacy of CFGs, even if the string set as a whole may be strictly speaking regular. Note that this approach holds that it is impossible to determine with any confidence that a particular string qua string is ungrammatical, but that it may be possible to tell one construction from another, and that the latter—and not the former—is the real basis of all linguistic work, theoretical, computational, and descriptive.

Finite Copying

The counterexamples to context-freeness in the literature have all been claimed to crucially involve expressions of unbounded length. This seemed necessary in view of the fact that an upper bound on length would imply finiteness of the subset of strings involved, which would as a result be of no formal language theoretic interest. However, it is often difficult to make a case for unbounded length, and the main result has been that, even though every linguist knows about reduplication, it seemed nearly impossible to find an instance of reduplication that could be used to make a formal argument against CFGs, even though no one would ever use a CFG to describe reduplication.

For, in addition to reduplications that can apply to unboundedly long expressions, there is a much better known class of reduplications exemplified by Indonesian pluralization of nouns. Here it is difficult to show that the reduplicated forms are infinite in number, because compound nouns are not pluralized in the same way, and ignoring compounding, it would seem that the number of nouns is finite. However, this number is very large and moreover it is probably not well defined. The class of noun stems is open, and can be enriched by borrowing from foreign languages and neologisms, and all of these spontaneously pluralize by reduplication.

Rounds, Manaster-Ramer, and Friedman (to appear) argue that facts like this mean that a natural language should not be modeled as a formal language but rather as a family of languages, each of which may be taken as an approximation to an ideal language. In the case before us, we could argue that each of the approximations has only a finite number of nouns, for example, but a different number in different approximations. This idea, related to the work of Yuri Gurevich on finite dynamic models of computation, allows us to state the argument that the existence of an open class of reduplications is sufficient to show the inadequacy of CFGs for that family of approximations. The basis of the argument is the observation that while each of the approximate languages could in principle have a CFG, each such CFG would differ from the next not only in the addition of a new lexical item but also in the addition of a new reduplication rule (for that particular item).

To capture what is really going on, we require a grammar that is the same for each approximation modulo the lexicon. This grammar in a sense generates the infinite ideal, but actually each actual approximate grammar only has a finite lexicon and hence actually only generates a finite number of reduplications. In order to model the flexibility of the natural language vocabulary, we assume that each member of the family has the same grammar modulo the terminal vocabulary and the rules which insert terminals.

Another way of stating this is that the lexicon of Indonesian is finite but of an indefinite size (what Gurevich calls "uncountably finite"). A CFG would still have to contain a separate rule for the plural of every noun and henc would have to be of an indefinite size. Thus, with addition of a new noun, the grammar would have to add a new rule. However, this would mean that the grammar at any given time can only form the plurals of nouns that have already been learned. Since speakers of the language know in advance how to pluralize unfamiliar nouns, this cannot be true. Rather the grammar at any given time must be able to form plurals of nouns that have not yet been learned. This in turn means that an indefinite number of plurals can be formed by a grammar of a determinate finite size. Hence, in effect, the number of rules for plural formation must be smaller than the number of plural forms that can be generated, and this in turn means that there is no CFG of Indonesian.

This brings up a crucial issue, of which we are all presumably aware but which is usually lost sight of in practice, namely, that the way a mathematical model (in this case, formal language theory) is applied to a physical or mental domain (in this case, natural language) is a matter of utility and not itself subject to proof or disproof. Formal language theory deals with sets of strings over well-defined finite vocabularies (also often called alphabets) such as the hackneyed {a, b}. It has been all too easy to fall into the trap of equating the formal language theoretic notion of vocabulary (alphabet) with the linguistic notion of vocabulary and likewise to confuse the formal language theoretic notion of a string (word) over the vocabulary (alphabet) with the linguistic notion of sentence.

However, the fundamental fact about all known natural languages is the openness of at least some classes of words (e.g., nouns but perhaps not prepositions or, in some languages, verbs), which can acquire new members through borrowing or through various processes of new formation, many of them apparently not rule-governed, and which can also lose members, as words are forgotten. Thus, the well-defined finite vocabularies of formal language theory are not a very good model of the vocabularies of natural languages. Whether we decide to introduce the notion of families of languages or that of uncountably finite sets or whether we rather choose to say that the vocabulary of a natural language is really infinite (being the set of all strings over the sounds or letters of the language that could conceivably be or become lexical items in it), we end up having to conclude that any language which productively reduplicates some open word class to form some grammatical category cannot have a CFG.

Copying in English

It should now be noted that reduplications (and reiterations generally) are extremely common in natural languages. Just how common follows from an inspection of the bewildering variety of such constructions that are found in English. All the examples cited here are productive though they may be of bounded length.

Linguistics shminguistics.

Linguistics or no linguistics, (I am going home).

A dog is a dog is a dog.

Philosophize while the philosophizing is good!

Moral is as moral does.

Is she beautiful or is she beautiful?

These are clause-level constructions, but we also find ones restricted to the phrase level.

(He) deliberates, deliberates, deliberates (all day long).

(He worked slowly) theorem by theorem.

(They form) a church within a church.

(He debunks) theory after theory.

Also relevant are cases where a copying dependency extends across sentence boundaries, as in discourses like:

A: She is fat.

B: She is fat, my foot.

It is interesting that several of these types are productive even though they appear to be based on what originally must have been more restricted, idiomatic expressions. The pattern *a* **X** *within a* **X**, for example, is surely derived from the single example *a state within a state*, yet has become quite productive.

Many of these patterns have analogues in other languages. For example, the **X** *after* **X** construction appears to involve quantification and this may be related to the fact that, for example, Bambara uses reduplication to mean 'whatever' and Sanskrit to mean 'every' (Pāṇini 8.1.4). English re**shm**uplication has close analogues in many languages, including the whole Dravidian and Turkic language families. Tamil **ki**duplication (e.g. *pustakam kistakam*) and Turkish **m**eduplication (e.g., *kitap mitap*) are instances of this, though the semantic range is somewhat different. In both of these, the sense is more like that of English *books and things*, *books and such*, i.e., a combination of deprecation and etceteraness rather than the purely derisive function of English *books shmooks*. The English **X** *or no* **X** ... pattern is very similar to a Polish construction consisting of the form **X** (nominative) **X** (instrumental) ... in its range of applications. The repetition of a verb or verbal phrase to deprecate excessive repetition or intensity of an action seems to be found in many languages as well.

I have not tried here to survey the uses to which copying constructions are put in different languages or even to document fully their wide incidence, though the examples cited should give some indication of both. It does appear that copying constructions are extremely common and pervasive, and this in turn suggests that they are central to man's linguistic faculties. When we consider such additional facts as the frequency of copying in child language, we may be tempted to take copying as one of the basic linguistic operations.

Copies vs. mirror images

The existence and the centrality of copying constructions poses interesting questions that go beyond the inadequacy of CFGs. For example, why should natural languages have reduplications when they lack mirror-image constructions, which are context-free? This asymmetry (first noted in Manaster-Ramer and Kac, 1985, and Rounds, Manaster-Ramer, and Friedman op. cit.) argues that it is not enough to make a small concession to context-sensitivity, as the saying goes. Rather than grudgingly clambering up the Chomsky Hierarchy towards Context-sensitive Grammars, we should consider going back down to Regular Grammars and striking out in a different direction. The simplest alternative proposal is a class of grammars which intuitively have the same relation to queues that CFGs have to stacks. The idea, which I owe to Michael Kac, would be that human linguistic processes make little if any use of stacks and employ queues instead.

Queue Grammars

This suggests that CFGs are not just inadequate as models of natural languages but inadequate in a particularly damaging way. They are not even the right point of departure, since they not only undergenerate but also overgenerate. This leads to the idea of a hierarchy of grammars whose relation to queues is like that of the Chomsky Grammars to stacks. A queue-based analogue to CFG is being developed, under the name of **Context-free Queue Grammar**. The current version is allowed rules of the following form:

A —> a

A —> aB

A —> aB...b

A —> a...b

A —> ...B

Whatever appears to the right of the three dots is put at the end of the string being rewritten. Otherwise, all definitions are as in a corresponding restricted CFG. Thus, the grammar

S —> aS...a

S —> bS...b

S —> a...a

S —> b...b

will generate the copying language over {a,b} excluding the null string and define derivations like the following:

S —> aSa —> abSab —> abaaba

S —> bSb —> baSba —> baaSbaa —> baabSbaab

On the other hand, I conjecture that the corresponding *xmi(x)* language cannot be generated by such a grammar. Even at this early stage of inquiry into these formalisms, then, we have some tangible promise of being able to explain why natural languages should have reduplications but not mirror-image constructions. Various *xh(x)* constructions such as the *respectively* ones and the cross-serial verb constructions can be handled in the same way as reduplications.

While the idea of taking queues as opposed to stacks as the principal nonfinite-state resource available to human linguistic processes would explain the prevalence of copying and the absence of mirror images, it does not explain the coexistence of center-embedded constructions with cross-serial ones or the relative scarcity of cross-serial constructions other than copying ones.

For this reason, if for no other, the CFQGs could not be an adequate model of natural language. In fact, there are further problems with these grammars. One way in which they fail is that they apparently can only generate two copies—or two cross-serially dependent substrings—whereas natural languages seem to allow more (as in *Grammar is grammar is grammar*). This is similar to the limitation of Head Grammars and Tree Adjoining Grammars to generating no more than four copies (Manaster-Ramer to appear a). However, a more general class of Queue Grammars appears to be within reach which will generate an arbitrary number of copies.

Perhaps more serious is the fact that CFQGs apparently can only generate copying constructions at the cost of profligacy (as defined in Rounds, Manaster-Ramer, and Friedman, to appear). The repair of this defect is less obvious, but it appears that the fundamental idea of basing models of natural languages on queues rather than stacks is not undermined. Rather, what is at issue is the way in which information is entered into and retrieved from the queue. The CFQGs suggest a piecemeal process but the considerations cited here seem to argue for a global one. A number of formalisms with these properties are being explored.

On the other hand, it may be that something much like the simple CFQG is a natural way of capturing cross-serial dependencies in cases other than copying. To see exactly what is involved, consider the difference between copying and other cross-serial dependencies. This difference has little to do with the form of the strings. Rather, in the case of other cross-serial dependencies, there is a syntactic and semantic relation between the nth elements of two or more structures. For example, in a *respectively* construction involving a conjoined subject and a conjoined predicate, each conjunct of the former is semantically combined with the corresponding conjunct of the latter. In the case of copying constructions, there is nothing analogous. The corresponding parts of the two copies do not bear any relations to each other. Thus it makes some sense to build up the corresponding parts of cross-serial construction in a piecemeal fashion, but this appears to be inapplicable in the case of copying constructions.

In view of all these limitations, the CFQGs might seem to be a non-starter. However, their importance lies in the fact that they are the first step in reorienting our notions of the formal space for models of natural language. Any real success in the theoretical models of human language depends on the development of appropriate mathematical concepts and on closing the gap between formal language and natural language theory. One of the first steps in this direction must involve breaking the spell of CFGs and the Chomsky Hierarchy. The CFQGs seem to be cut out for this task. Moreover, the idea that queues rather than stacks are involved in human language appears to be correct, and this more general result is independent of the limitations of CFQGs. However, given my stated goals for formal models, it is necessary to develop models such as CFQGs before proceeding to more complex ones precisely in order to develop an appropriate notion of formal space within which we will have to work.

The other main point addressed in this paper, the need to model human languages as families of formal languages or as formal languages with indefinite terminal vocabularies, is intended in the same spirit. The allure of identifying formal language theoretic concepts with linguistic ones in the simplest possible way is hard to overcome, but it must be if we are to get any meaningful results about natural languages through the formal route. It will, again, be necessary to do more work on these concepts, but it is beginning to look as though we have found the right direction.

REFERENCES

Carlson, Greg N. 1983. Marking Constituents. **Linguistic Categories** (Frank Heny and Barry Richards, eds.), **1**: Categories, 69-98. Dordrecht: Reidel.

Chomsky, Noam. 1963. Formal Properties of Grammars. **Handbook of Mathematical Psychology** (R. Duncan Luce at al., eds.), **2**: 323-418. New York: Wiley.

Culy, Christopher. 1985. The Complexity of the Vocabulary of Bambara. **Linguistics and Philosophy, 8**: 345-351.

Daly, R. T. 1974. **Applications of the Mathematical Theory of Linguistics**. The Hague: Mouton.

Gazdar, Gerald, and Geoffrey K. Pullum. 1985. Computationally Relevant Properties of Natural Languages and Their Grammars. **New Generation Computing, 3**: 273-306.

Higginbotham, James. 1984. English is not a Context-free Language. **Linguistic Inquiry, 15**: 225-234.

Kac, Michael B. To appear. Surface Transitivity and Context-freeness.

Kac, Michael B., Alexis Manaster-Ramer, and William C. Rounds. To appear. Simultaneous-distributive Coordination and Context-freeness. **Computational Linguistics.**

Langendoen, D. Terence. 1977. On the Inadequacy of Type-3 and Type-2 Grammars for Human Languages. **Studies in Descriptive and Historical Linguistics: Festschrift for Winfred P. Lehmann** (Paul Hopper, ed.), 159-171. Amsterdam: Benjamins.

Langendoen, D. Terence, and Paul M. Postal. 1984. Comments on Pullum's Criticisms. **CL, 8**: 187-188.

Levelt, W. J. M. 1974. **Formal Grammars in Linguistics and Psycholinguistics**. The Hague: Mouton.

Manaster-Ramer, Alexis. 1983. The Soft Formal Underbelly of Theoretical Syntax. **CLS, 19**: 256-262.

Manaster-Ramer, Alexis. To appear a. Dutch as a Formal Language. **Linguistics and Philosophy**.

Manaster-Ramer, Alexis. To appear b. Subject-verb Agreement in Respective Coordinations in English.

Manaster-Ramer, Alexis, and Michael B. Kac. 1985. Formal Languages and Linguistic Universals. Paper read at the Milwaukee Symposium on Typology and Universals.

Postal, Paul M. 1964. Limitations of Phrase Structure Grammars. **The Structure of Language: Readings in the Philosophy of Language** (Jerry A. Fodor and Jerrold J. Katz, eds.), 137-151. Englewood Cliffs, NJ: Prentice-Hall.

Postal, Paul M., and D. Terence Langendoen. 1984. English and the Class of Context-free Languages. **CL, 10**:177-181.

Pullum, Geoffrey K., and Gerald Gazdar. 1982. Natural Languages and Context-free Languages. **Linguistics and Philosophy, 4**: 471-504.

Pullum, Geoffrey K. 1984a. On Two Recent Attempts to Show that English is not a CFL. **CL, 10**: 182-186.

Pullum, Geoffrey K. 1984b. Syntactic and Semantic Parsability. **Proceedings of COLING84**, 112-122. Stanford, CA: ACL.

Rounds, William C., Alexis Manaster-Ramer, and Joyce Friedman. To appear. Finding Natural Languages a Home in Formal Language Theory. **Mathematics of Language** (Alexis Manaster-Ramer, ed.). Amsterdam: John Benjamins.

Shieber, Stuart M. 1985. Evidence against the Context-freeness of Natural Language. **Linguistics and Philosophy, 8**: 333-343.