CHRISTOPHER CULY*

THE COMPLEXITY OF THE VOCABULARY OF BAMBARA

In this paper I look at the possibility of considering the vocabulary of a natural language as a sort of language itself. In particular, I study the weak generative capacity of the vocabulary of Bambara, and show that the vocabulary is not context free. This result has important ramifications for the theory of syntax of natural language.

A language can be defined, from the point of view of formal language theory, as being "a set of strings of symbols from some one alphabet (Hopcroft and Ullman, 1979, p. 2), where a string is "a finite sequence of symbols juxtaposed" (Hopcroft and Ullman, 1979, p. 1), and an alphabet is a "a finite set of symbols" (Hopcroft and Ullman, 1979, p. 2).[1] Given a language, one can study its complexity in different ways. The weak generative capacity of a language is the complexity of the set of strings of the language. The strong generative capacity of a language is the complexity of the set of structures that are assigned to the strings of the language.

In terms of generative capacity, linguists usually think of the case where the "alphabet" is the vocabulary of a natural language, and the "strings of symbols" are strings of vocabulary items, i.e., sentences. There has been a lot of contróversy concerning the generative capacity, taken in this way, of natural language. I will not go into details here, but see Pullum and Gazdar (1982) for a lengthy discussion, and Bresnan et al. (1982) and Culy (1983) for more recent developments.

Returning to the definition of language, this time considering the vocabulary of a natural language, we see that the vocabulary itself can be thought of as a language in the above sense. In this case, the "alphabet" is the set of morphemes of the natural language, and the "strings of symbols" are strings of morphemes. Given this observation that the vocabulary of a natural language is itself a language, we can study the weak and strong generative capacities of the vocabulary. For the rest of the paper, I consider the weak generative capacity of the vocabulary of Bambara, a Northwestern Mande language spoken in Mali and neighboring countries.

Bambara has a construction of the form Noun *o* Noun, where the two nouns have the same form. This construction translates as "whichever Noun" or "whatever Noun".[2]

Linguistics and Philosophy **8** (1985) 345–351. 0165–0157/85.10

(1)(a) wulu o wulu
dog dog

"whichever dog"

(b) malo o malo
uncooked rice rice

"whatever uncooked rice"

(c) *wulu o malo[3]
dog rice

(d) *malo o wulu
rice dog.

This construction is very productive, with few, if any, restrictions on the choice of the noun.

There is evidence that the Noun *o* Noun construction belongs in the vocabulary rather than in the syntax. Bambara is a tone language, and as such it has two types of rules governing the interaction of tones: rules dealing with the interaction of adjacent lexical items, and rules dealing with the interaction of components of a compound, be it nominal, verbal, or whatever. Internally, the Noun *o* Noun construction does not follow the rules for adjacent lexical items, but rather has its own peculiar rule. (Cf. Bird et al., pp. 8–9, 166, for a description of the first sort of rules and for the Noun *o* Noun construction.) Thus, tonal evidence indicates that the Noun *o* Noun construction does indeed belong in the vocabulary rather than the syntax.

Bambara also has an agentive construction: Noun(N) + Transitive Verb(TV) + *la*, which translates as "one who TVs Ns".

(2)(a) wulu + nyini + la = wulunyinina[4]
dog search for

"one who searches for dogs", i.e., "dog searcher"

(b) wulu + filè + la = wulufilèla
dog watch

"one who watches dogs", i.e., "dog watcher"

(c) malo + nyini + la = malonyinina
rice search for

"one who searches for rice", i.e., "rice searcher"

(d) malo + filè + la = malofilèla
rice watch

"one who watches rice", i.e., "rice watcher".

This construction is also very productive, with interpretability being virtually the only restriction. In particular, the construction is recursive, that is, the noun in the construction can be of the same form.[5]

(3)(a) wulunyinina + nyini + la = wulunyininanyinina
dog searcher search for

"one who searches for dog searchers"

(b) wulunyinina + filè + la = wulunyininafilèla
dog searcher watch

"one who watches dog searchers"

(c) wulufilèla + nyini + la = wulufilèlanyinina
dog watcher search for

"one who searches for dog watchers"

(d) wulufilèla + filè + la = wulufilèlafilèla
dog watcher watch

"one who watches dog watchers"

(e) malonyinina + nyini + la = malonyininanyinina
rice searcher search for

"one who searches for rice searchers"

(f) malonyinina + filè + la = malonyininafilèla
rice searcher watch

"one who watches rice searchers"

(g) malofilèla + nyini + la = malofilèlanyinina
rice watcher search for

"one who searches for rice watchers"

(h) malofilèla + filè + la = malofilèlafilèla
rice watcher watch

"one who watches rice watchers".

These agentive nouns from the second construction can be used in the Noun *o* Noun construction.

(4)(a) wulunyinina o wulunyinina
dog searcher dog searcher

"whichever dog searcher"

(b) wulufilèla o wulufilèla
dog watcher dog watcher

"whichever dog watcher"

(c) wulunyininanyinina o wulunyininanyinina
one who searches for dog searchers
one who searches for dog searchers

"whoever searches for dog searchers"

(d) wulunyininafilèla o wulunyininafilèla
one who watches dog searchers
one who watches dog searchers

"whoever watches dog searchers"

(e) wulufilèlanyinina o wulufilèlanyinina
one who searches for dog watchers
one who searches for dog watchers

"whoever searches for dog watchers"

(f) wulufilèlafilèla o wulufilèlafilèla
one who watches dog watchers
one who watches dog watchers

"whoever watches dog watchers"

(g) malonyinina o malonyinina
rice searcher rice searcher

"whichever rice searcher"

(h) malofilèla o malofilèla
rice watcher rice watcher

"whichever rice watcher"

(i) malonyininanyinina o malonyininanyinina
one who searches for rice searchers
one who searches for rice searchers

"whoever searches for rice searchers"

(j) malonyininafilèla o malonyininafilèla
one who watches rice searchers
one who watches rice searchers

"whoever who watches rice searchers"

(k) malofilèlanyinina o malofilèlanyinina
one who searches for rice watchers
one who searches for rice watchers

"whoever searches for rice watchers"

(l) malofilèlafilèla o malofilèlafilèla
one who watches for rice watchers
one who watches for rice watchers

"whoever watches rice watchers".

The two nouns still have to have the same form.

(5)(a) *wulunyinina o wulufilèla
dog searcher dog watcher

(b) *wulunyinina o malonyinina
dog searcher rice searcher

(c) *wulunyinina o malofilèla
dog searcher rice watcher.

This very free process of redoubling causes the vocabulary of Bambara to be non-context-free[6,7] as I now show. Let B be the vocabulary of Bambara. Thus, B is a set of strings of morphemes. Let

$$R = \{\text{wulu(filèla)}^h\text{(nyinina)}^i \text{ o wulu(filèla)}^j\text{(nyinina)}^k \mid h, i, j, k \geq 1\}.$$

The intersection of B and R is

$$B \cap R = B' = \{\text{wulu(filèla)}^m\text{(nyinina)}^n \text{ o wulu(filèla)}^m\text{(nyinina)}^n \mid m, n \geq 1\}.$$

B' is of the form $\{a^m b^n a^m b^n \mid m, n \geq 1\}$ (the *o* can be disregarded without loss of generality), and hence, it is easy to show that it is not context-free (cf. Hopcroft and Ullman, 1979, p. 136, Example 6.5). Since R is a regular language, and the intersection of a context-free language and a regular language is always a context-free language (cf. Hopcroft and Ullman, 1979, p. 135), if B were context-free, B' would also be context-free. But B' is not context-free, so neither is B. Thus, the vocabulary of Bambara is not context-free.

Note that the above argument does not determine how complex the vocabulary of Bambara is. It merely gives a lower bound for the weak generative capacity. Hence, one is led to the conclusion that the complexity of the vocabulary of a natural language can be more than context-free.

This argument raises several interesting points. The first point is, can one find a smallest upper bound for the complexity of the vocabulary of a natural language? One can also divide the vocabulary into subsets and consider the generative capacity of each subset. For example, in the above case of Bambara, B' was actually just a set of nouns. One could also consider verbs, adjectives, etc.

The other points have to do with syntax. For Bambara at least, and probably for many other languages (cf. Pullum and Gazdar, 1982; Langendoen, 1981; Carden, 1983) there are an infinite number of vocabulary items from which sentences can be formed. This is contrary to the definition of language given at the beginning since the "alphabet" is no longer finite. We can get around this point by saying that the syntax generates a language of strings of lexical categories (with all their features). The natural language is obtained by substituting items from the vocabulary for the lexical categories (which is how we tend to think of things, intuitively at least). That is, a natural language can be obtained by the composition of two other languages.

It turns out that once we allow the vocabulary of a language to be infinite, we have to have something like the substitution mentioned above if we want to keep the syntax in a reasonable state. Each symbol in the syntax must be introduced by some rule, so if we have an infinite number of vocabulary items to be introduced in the syntax, we have to have an infinite number of rules. This is a highly undesirable state of affairs from the point of view of the weak generative capacity of the syntax (cf. Culy, 1982; Peters and Uszkoreit, 1982). Thus, syntacticians studying weak generative capacity should consider the language to be strings of lexical categories rather than vocabulary items. Since the natural language is obtained by substitution, they do not have to worry about actual vocabulary items, just the lexical categories. That is to say, the study of the generative capacity of the syntax is independent of the study of the generative capacity of the vocabulary.

Notes

* My sincere appreciation goes to Adama Koné and Saloum Soumaré and especially to the Center for the Study of Language and Information (CSLI), Stuart Shieber, and Thomas Wasow for their generous help in preparing this article. Of course, all deficiencies are the responsibility of the author.

[1] This is only one narrow point of view which ignores many aspects of language, including, among other things, the meanings associated with the symbols.

[2] The Bambara is transcribed in the official Malian orthography.

[3] I want to thank the two anonymous referees for pointing out the necessity of including the ungrammatical examples.

[4] Due to a very pervasive rule, /l/ becomes [n] after a syllable containing a nasal consonant or nasal vowel.

[5] These constructions soon become awkward, due to their length. However, my informants maintained the grammaticality of the examples.

[6] The Chomsky hierarchy is a means of classifying the complexity of languages. There are four successive levels, each level properly including the ones before it. The least complex level is that of regular languages, followed by context-free, context-sensitive, and finally recursively enumerable languages.

[7] This is in answer to Langendoen (1981), who states he knows of no language with a vocabulary more complex than a regular language.

References

Bailleul, C.: 1981, *Petit Dictionnaire Bambara-Français Français-Bambara*, Avebury Publishing Company, England.

Bird, C., J. Hutchison, and M. Kanté: 1977, *An Ka Bamanankan Kalan: Beginning Bambara*, Indiana University Linguistics Club, Bloomington, Indiana.

Bresnan, J., R. M. Kaplan, S. Peters, and A. Zaenen: 1982, 'Cross-serial Dependencies in Dutch', *Linguistic Inquiry* **13**, 613–635.

Carden, G.: 1983, 'The non-finiteness of the word formation component', *Linguistic Inquiry* **14**, 537–547.

Culy, C.: 1982, 'String Variables and Metarules', unpublished manuscript, Department of Linguistics, Stanford University, Stanford, California.

Culy, C.: 1983, *An Extension of Phrase Structure Rules and Its Application to Natural Language*, unpublished M.A. thesis, Stanford University, Stanford, California.

Hopcroft, J. E. and J. D. Ullman: 1979, *Introduction to Automata Theory, Languages, and Computation*, Addison-Wesley, Reading, Massachusetts.

Langendoen, T.: 1981, 'The Generative Capacity of Word-Formation Components', *Linguistic Inquiry* **12**, 320–322.

Peters, P. S., and H. Uszkoreit: 1982, 'Essential Variables in Meta-rules', paper presented at the annual meeting of the Linguistic Society of America, San Diego, December.

Pullum, G. K. and G. Gazdar: 1982, 'Natural Languages and Context-free Languages', *Linguistics and Philosophy* **4**, 471–504.

Peace Corps
Bamako
Mali