Précis of Kobele [2006]

Gregory M. Kobele

This dissertation makes three main contributions. First, I show how to assign meanings (model-theoretic objects) to sentences in a way that takes advantage of the inductive structure minimalist grammars assign to the sentences they generate.[1] In other words, I show how to associate a compositional semantics with minimalist grammars. To show that this is more than a mathematical exercise, I provide grammar fragments for English and the west african language Yoruba, and show how the developed semantics assigns reasonable truth conditions to the expressions of these fragments.

Second, I present a formalization of the recently influential copy theory of movement [Chomsky, 1995, Nunes, 2004], and show that such an extension to the minimalist grammar formalism does not increase its expressive power too much (i.e. all languages definable by minimalist grammars with copying are parallel multiple context-free languages (PMCFLs [Seki et al., 1991]), and are thus strictly contained in **P**, the languages recognizable in polynomial time on a deterministic turing machine).

Finally, using original data elicited from a native speaker of Yoruba, I argue that natural languages, regardless of the syntactic theory one uses to describe them, require more in the way of abstract computational resources for their description than standardly assumed [Joshi, 1985]. Thus I motivate the increase in expressive power obtained by adding copying to minimalist grammars, by arguing that no grammar formalism without the ability to describe languages like $\{a^{2^n} : n \in \mathbb{N}\}$ (a language generable by minimalist grammars with copying but not without) is empirically adequate.

This final contribution is the most significant, telling us the most about the phenomenon of natural language. For this reason, I focus solely upon it in the space remaining.

[1] Minimalist grammars are introduced in [Stabler, 1997], and are intended to provide a formalization of the core ideas in the minimalist program [Chomsky, 1995].

1 On the complexity of natural language

In final part of the dissertation, I step back from any particular theory of language, and focus rather on the shape and complexity of the constructions that manifest themselves. Examining the pattern of copying present in the Yoruba relativized predicate construction, I conclude that any satisfactory grammar for Yoruba must be capable of copying copies (of copies...), and thus also of generating languages like a^{2^n} (the language where each sentence consists of, for some number n, exactly 2^n as).

This conclusion is a momentous one, contradicting (if not in letter then in spirit) the widely held hypothesis of the mild context sensitivity of natural language, which we recount in § 1.1. As a practical consequence, it means that many of our grammatical theories are too weak to describe natural languages, predicting of things that do in fact exist, that they couldn't possibly.[2]

These conclusions have been argued for before. Perhaps the best known arguments are Becker et al.'s [1992] based on scrambling in German, and Michaelis and Kracht's [1997] on Suffixaufnahme (case stacking) in Old Georgian. Neither of these arguments have been decisively refuted, although their initial force has been tempered greatly via a reanalysis of the motivating data [Joshi et al., 2000, Bhatt and Joshi, 2004].

There are two major differences between the present argument, and its predecessors. First is the obvious linguistic naturalness of the construction type investigated here, and the plethora of analyses of disparite phenomena that utilize mechanisms designed to capture this very type of construction. Second is the development of a constrained formalism (minimalist grammars with copy movement) which is able to efficiently assign the obvious structures to the constructions in question. Thus, although I conclude that natural languages require more than mild context sensitivity for their description, I hypothesize that they don't require much more than that.

[2]We can really only assess this for those theories that have been made explicit. Of these, two of the most widely used, Combinatory Categorial Grammar and Tree Adjoining Grammar in any of its incarnations (except those, such as explored in [Chen-Main, 2006], which add an operation of syntactic copying), are too weak. Head-Driven Phrase Structure Grammar, another popular theory, is untouched by this result, as the formalism makes no non-trivial predictions about natural language (see, e.g. Kracht [2003]).

1.1 Mild context-sensitivity

The hypothesis of mild context sensitivity, first articulated in [Joshi, 1985], is a claim that all human languages share certain characteristic properties, and that, moreover, these properties are non-accidental from a grammatical perspective. In other words, our theories of language should predict that only languages that are mildly context-sensitive exist. Like claims that natural language is regular or context-free, the claim that natural languages are mildly context-sensitive is statable independently of any particular grammar formalism, which fact makes its empirical content crystal clear.

Mild context-sensitivity can be given a bipartite characterization, as a list of conditions that a language must meet.[3] The first condition is that the language be among those whose strings are recognizable by a deterministic turing machine in polynomial time. This is often called *efficient recognizability*. Although it contains the words 'efficient' and 'recognize', this criterion is emphatically *not* related to ideas about human language processing. We might just as well have characterized this property as definability in first order logic with a least fixed point operator (see e.g. [Immerman, 1995]). The essence of this criterion is to circumscribe a class of patterns of reasonable complexity. It is a non-trivial property. Copying (ww), reversal (ww^r), and exponential (a^{2^n}) patterns are efficiently recognizable, whereas primacy ($\{a^p : p \text{ is prime}\}$) and theorem-hood ($\{a^g : g \text{ is the gödel number of a theorem of FOL}\}$) are not. The second condition is that the language be of *constant growth*.[4] A language is of constant growth just in case the size of its strings doesn't grow 'too quickly'. Intuitively, the idea is that at each step in the derivation of a sentence, the rules add only a fixed amount of new material. The language a^{2^n} is not of constant growth. Constant growth says nothing about how the words within a sentence are arranged, and is therefore independent of the criterion of efficient recognizability.[5]

[3] A third condition is often added to this list. This third condition requires that there be a limited number of cross-serial dependencies. This intuition has been notoriously difficult to pin down in a meaningful, grammar independent, way. As it is not clear what it should mean, I leave it out of my characterization of mild context-sensitivity.

[4] A stronger condition, more in keeping with the intent of the 'constant growth' criterion, is *semilinearity*. Whereas constant growth requires that the lengths of sentences not grow too fast, semilinearity requires this of the numbers of the individual words in a sentence. The language $a^{2^n}b^*$ (the language which has an exponential number of as followed by any number of bs) is of constant growth, but is not semilinear.

[5] Although the language of primes, a^p is not efficiently recognizable, and thus neither

1.2 The structure of a challenge

At first blush it seems easy enough to mount a challenge to the MCS hypothesis. We might try and look for a natural language that wasn't recognizable in polynomial time or wasn't of constant growth. However, there is no way to challenge the MCS hypothesis on the basis of a corpus of data. The MCS hypothesis rules out only certain infinite sets of sentences. Corpora are of necessity finite. In order to challenge the MCS hypothesis, we need to first generalize from the observed data to an infinite set of potential data, of which the observed data is but a small sample. Then we can decide whether or not our generalization is compatible with the MCS hypothesis.

A good challenge will have the following two attributes. First, the generalization argued for is a reasonable one, in the sense that it is an instance of an already established type. That is, for the proposal to be empirically secure, we expect it to be parsimonious, compatible with and even supported by other independently motivated assumptions about language mechanisms. Second, there aren't alternative reasonable generalizations compatible with the data that are compatible with the MCS hypothesis. Accordingly, a response to a good challenge takes the form of coming up with a novel reasonable generalization compatible with both the data and the MCS hypothesis.

1.3 Yoruba

Here I will outline the challenge to the MCS hypothesis presented in the dissertation, based on the relativized predicate construction in Yoruba. I will argue that Yoruba relativized predicates can themselves contain relativized predicates (which may in turn contain relativized predicates etc.). This means that copies can be of copies. In other words, copying operations can apply iteratively. Although the set of sentences of Yoruba does not in itself cause trouble for the MCS hypothesis, the mechanisms that we require to describe these sentences elegantly do. The claim is, then, that as soon as we are able to give a natural account of languages like Yoruba, we are also able

is the language $a^p b^*$ (the language with a prime number of as followed by any number of bs), this latter *is* of constant growth. In other words, we can 'pad out' a set of strings of non-constant growth with dummy symbols to satisfy the constant growth property. The stronger property, semilinearity, is also independent of efficient recognizability, with $a^p b^* + b^* a^*$ (the language in which as precede bs if there are a prime number of them, and follow the bs otherwise) being semilinear, but not efficiently recognizable.

to describe non-MCS languages using the *very same* mechanisms in the *very same* way. Therefore, while it may be true that all attested languages are in fact semilinear, this fact, like the fact that ww^r (the language of palindromes) is not an attested language, cannot receive a syntactic explanation.

My argument is simple and can be summarized as follows. I begin by arguing that the relativized predicate construction in Yoruba involves copying (1.3.1). Next I will argue that relative clauses can be copied in the relativized predicate construction (1.3.2). As relative clauses are themselves clauses, and can contain arguments beyond the one abstracted over, we must countentance one of these arguments being itself a relativized predicate, and thus that relativized predicates may be copied in the relativized predicate construction, from which the broader conclusion follows.

1.3.1 Relativized predicates

Verb phrases may appear in constructions which look very much like relative clauses. When they do, the resulting phrase has the distribution of a(n abstract) noun. Of particular interest to us is the relation between the verb (phrase) which acts as head of the relative clause, and the verb (phrase) that is inside the relative clause. Consider the following examples. 1 is a simple transitive sentence. In 2, a (nominalized) copy of the verb heads the relativized predicate. We see that it must be a copy of the verb, as evidenced by the ungrammaticality of 3.[6]

(1) *Jimọ ra adiẹ*
Jimọ buy chicken
"Jimọ bought a chicken."

(2) *Rira ti Jimọ ra adiẹ*
buying TI Jimọ buy chicken
"The fact/way Jimọ bought a chicken"

(3) **Jije ti Jimọ ra adiẹ*
eating TI Jimọ buy chicken

In example 4, a (nominalized) copy of the verb phrase *ra adiẹ* acts as the head of the relativized predicate. Again, we see that it must be a copy of

[6]Yoruba orthography uses the letters ọ and ẹ to represent the same sounds as IPA ɔ and ɛ respectively. Yoruba also has three level tones, which I do not represent here.

the verb phrase, as changing the verb (as in 5) or the object (as in 6) lead to ungrammaticality.

(4) *Rira adiẹ ti Jimọ ra adiẹ*
buying chicken TI Jimọ buy chicken
"The fact/way Jimọ bought a chicken"

(5) **Jije adiẹ ti Jimọ ra adiẹ*
eating chicken TI Jimọ buy chicken

(6) **Rira nkan ti Jimọ ra adiẹ*
buying something TI Jimọ buy chicken

The obvious generalization is that in a verbal relative clause the element to the left of *ti* is a copy of the (main) predicate to the right of the *ti*. Somewhat schematically, we might represent the possible relativized predicates given a sentence of type S V O in the following terms.

V (O) *ti* S V O

1.3.2 On the size of the copied object

Having established that the relativized predicate construction does indeed involve copying, we turn now to the question of how much material can be copied. Whatever mechanism we decide to use to account for the Yoruba data, it must be able to copy *unboundedly much* material. We base this on the fact (to be shown) that the copied object is in fact a DP, as evidenced by the fact that it may head a relative clause. As DPs are unbounded in size (as one may always add yet another relative clauses), the conclusion follows.

The examples below show that relative clauses can be copied in the relativized predicate construction. Sentence 7 shows our simple SVO sentence 1 with the addition of a relative clause modifying the object *adiẹ*.

(7) *Jimọ ra [adiẹ ti o gọ]*
Jimọ buy chicken TI 3S dumb
"Jimọ bought the stupid chicken."

The remaining examples show that although a relative clause can appear either just in the lower copy (8), just in the higher copy (9), or in both (10), if there are relative clauses in both higher and lower VPs, they must be identical (11). In other words, although relative clauses are not *required* to be copied, the grammar of Yoruba *allows* them to be.

(8) *Rira adiẹ ti Jimọ ra [adiẹ ti o gọ]*
buying chicken TI Jimọ buy chicken TI 3S dumb

(9) *Rira [adiẹ ti o gọ] ti Jimọ ra adiẹ*
buying chicken TI 3S dumb TI Jimọ buy chicken

(10) *Rira [adiẹ ti o gọ] ti Jimọ ra [adiẹ ti o gọ]*
buying chicken TI 3S dumb TI Jimọ buy chicken TI 3S dumb

(11) **Rira [adiẹ ti o gọ] ti Jimọ ra [adiẹ ti o kere]*
buying chicken TI 3S dumb TI Jimọ buy chicken TI 3S small

1.4 Evaluating the Challenge

The challenge to the MCS hypothesis is that given that Yoruba has constructions which involve copying of arbitrarily large structures which may themselves contain copied structures, we need to have on hand a mechanism that is able to copy copies. Once we have such a mechanism in our grammar, we are able to generate languages that are not of constant growth.

Our challenge is a strong one. Not only is the generalization we have argued for the obvious one, it is robustly attested in numerous West African (influenced) languages. Bùlì (Gur) [Hiraiwa, 2005], Krio (Creole) [Nylander, 1985], Twi (Kwa), Vata (Kru) [Koopman, 1983], Wolof (Atlantic) and many others have relativized predicate constructions, in which a copy of a verbal constitutent appears both as the head of the relative clause, and internally to the clause. These languages differ with respect to the size of the copied constituent, with Twi on one end of the spectrum allowing only a single verb root to be copied, Wolof somewhere in the middle permitting verbal complexes to be copied, and Yoruba on the other end allowing full VPs to be copied. Moreover, analyses appealing to mechanisms like ours abound in the literature. Copying of some sort has been assumed to underlie ellipsis [Sag, 1976, Chung et al., 1995, Murguía, 2004], A-not-A questions in Mandarin Chinese [Radzinski, 1990, Huang, 1991], predicate clefts in languages as diverse as Hebrew [Landau, 2006] and Korean [Cho and Nishiyama, 2000], and free relatives in languages as diverse as Bambara [Culy, 1985] and Italian

[Gullì, 2003], not to mention the plethora of analyses of non-surface-copying phenomena using the copy theory of movement in the minimalist program.

1.5 Conclusions

Copying is big trouble for current MCS formalisms—they can do it to a limited extent, but only by encoding the string component of a lexical item in its category. This makes the structures assigned to copies very unnatural, and therefore the *same* syntactic generalization needs to be stated twice: once over non-copies, and then once over copies, making the resulting grammars unnecessarily complex (a point made nicely by Pullum [2006]). The natural and obvious generalizations I have drawn about Yoruba, where there exist copies with copies (of copies...) contained in them, is impossible to be stated in these formalisms.

References

Tilman Becker, Owen Rambow, and Michael Niv. The derivational generative power of formal systems, or, scrambling is beyond LCFRS. Technical Report IRCS-92-38, Institute for Research in Cognitive Science, University of Pennsylvania, 1992.

Rajesh Bhatt and Aravind Joshi. Semilinearity is a syntactic invariant: A reply to Michaelis and Kracht (1997). *Linguistic Inquiry*, 35(4):683–692, Fall 2004.

Joan Chen-Main. *On the generation and linearization of multi-dominance structures*. PhD thesis, Johns Hopkins, 2006.

Eun Cho and Kunio Nishiyama. Yoruba predicate clefts from a comparative perspective. In Vicki Carstens and Frederick Parkinson, editors, *Advances in African Linguistics*, volume 4 of *Trends in African Linguistics*, pages 37–49. Africa World Press, Trenton, NJ, 2000.

Noam Chomsky. *The Minimalist Program*. MIT Press, Cambridge, Massachusetts, 1995.

Sandra Chung, William A. Ladusaw, and James McCloskey. Sluicing and logical form. *Natural Language Semantics*, 3(3):239–282, 1995.

Christopher Culy. The complexity of the vocabulary of Bambara. *Linguistics and Philosophy*, 8(3):345–352, 1985.

Antonino Gullì. *Phrasal Reduplication in Syntax*. PhD thesis, The City University of New York, 2003.

Ken Hiraiwa. *Dimensions of Symmetry in Syntax: Agreement and Clausal Architecture*. PhD thesis, Massachusetts Institute of Technology, 2005.

C.-T. James Huang. Modularity and Chinese A-not-A questions. In Carol Georgopoulos and Roberta Ishihara, editors, *Interdisciplinary Approaches to Language, Essays in Honor of S.-Y. Kuroda*, pages 305–332. Kluwer, Dordrecht, 1991.

Neil Immerman. Descriptive complexity: a logician's approach to computation. *Notices of the American Mathematical Society*, 42(10):1127–1133, 1995.

Aravind Joshi. How much context-sensitivity is necessary for characterizing structural descriptions. In David Dowty, Lauri Karttunen, and Arnold Zwicky, editors, *Natural Language Processing: Theoretical, Computational and Psychological Perspectives*, pages 206–250. Cambridge University Press, NY, 1985.

Aravind K. Joshi, Tilman Becker, and Owen Rambow. Complexity of scrambling: A new twist to the competence-performence distinction. In Anne Abeillé and Owen Rambow, editors, *Tree Adjoining Grammars: Formalisms, Linguistic Analysis and Processing*. CSLI Publications, Stanford, 2000.

Gregory M. Kobele. *Generating Copies: An investigation into structural identity in language and grammar*. PhD thesis, University of California, Los Angeles, 2006.

Hilda Koopman. *The Syntax of Verbs: From Verb Movement Rules in the Kru Languages to Universal Grammar*. Foris, Dordrecht, 1983.

Marcus Kracht. *The Mathematics of Language*, volume 63 of *Studies in Generative Grammar*. Mouton de Gruyter, Berlin, 2003.

Idan Landau. Chain resolution in Hebrew V(P)-fronting. *Syntax*, 9(1):32–66, April 2006.

Jens Michaelis and Marcus Kracht. Semilinearity as a syntactic invariant. In Christian Retoré, editor, *Logical Aspects of Computational Linguistics*, pages 37–40, NY, 1997. Springer-Verlag (Lecture Notes in Computer Science 1328).

Elixabete Murguía. *Syntactic Identity and Locality Restrictions on Verbal Ellipsis*. PhD thesis, University of Maryland, 2004.

Jairo Nunes. *Linearization of Chains and Sideward Movement*, volume 43 of *Linguistic Inquiry Monographs*. MIT Press, Cambridge, Massachusetts, 2004.

Dudley K. Nylander. Factivity, presupposition and the relativised predicate in Krio. *Studies in African Linguistics*, 16(3):323–336, December 1985.

Geoffrey K. Pullum. Render unto syntax the things which are syntax. handout of talk, September 2006. Available at `http://www.phon.ucl.ac.uk/research/pullum.pdf`.

Daniel Radzinski. Unbounded syntactic copying in Mandarin Chinese. *Linguistics and Philosophy*, 13(1):113–127, 1990.

Ivan A. Sag. *Deletion and Logical Form*. PhD thesis, Massachusetts Institute of Technology, Cambridge, Massachusetts, 1976.

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

Edward P. Stabler. Derivational minimalism. In Christian Retoré, editor, *Logical Aspects of Computational Linguistics*, volume 1328 of *Lecture Notes in Computer Science*, pages 68–95. Springer-Verlag, Berlin, 1997.