A NOTE ON xy LANGUAGES*

1. A CLAIM ABOUT xy LANGUAGES

Pullum and Gazdar (1982, hereafter PG), discussing the context-freeness of natural languages, include a section (pp. 476–9) in which they discuss Chomsky's (1963, pp. 378–9) claims about English comparatives and ‘xy languages’, those whose grammar requires nonidentity between substrings x and y. Chomsky maintained that no language in one class of xy languages (call this class \(\mathcal{L}\)) is context-free (CF); PG argue that \(\mathcal{L}\) includes an infinite set of xy languages that ARE CF.

We will show that though PG's point is well taken, it does not get at the substance of Chomsky's claim, because \(\mathcal{L}\) also includes an infinite set of xy languages that are NOT CF. As background to a demonstration of this point, we first define \(\mathcal{L}\).

A language belongs to \(\mathcal{L}\) if it consists of all strings of the form \(ax\beta y\gamma\), where

\[
\begin{align*}
(1)(a) & \quad \alpha, \beta, \text{ and } \gamma \text{ are fixed strings;} \\
(b) & \quad x \text{ and } y \text{ both belong to some base language } L_B; \\
(c) & \quad x \neq y; \\
(d) & \quad L_B \text{ is infinite; and} \\
(e) & \quad L_B \text{ is CF.}
\end{align*}
\]

Clause (a) allows for invariant portions of the sentences in the xy language. Clause (b) says that the variable portions are sentences in the base language. Clause (c) is the nonidentity requirement. Clause (d), given explicitly by Chomsky, is necessary to exclude finite base languages, since for them an xy language will also be finite. Clause (e), given neither by Chomsky nor PG, is necessary to eliminate non-CF base languages, for which an xy language is not necessarily CF.\(^1\)

PG's claim is supported by an illustrative example in which the base language is \((a + b)^*\), the set of strings on the alphabet \(\{a, b\}\): This base language is infinite, and it certainly is CF (indeed, it is regular). The point could have been made with an even simpler base language, namely \(a^*\). An xy language on this base is the set of all strings of the form \(aa^n\beta a^n\gamma\), where \(m \neq n\), i.e., the set of all strings of this form in which \(m\) is either less
(2) \[ S \to a \{ a \{ T_1 \} a \} \gamma \]

\[ T_1 \to a \{ T_1 \} (a) \]

\[ T_2 \to a \{ T_2 \} a \]

What is common to these two examples is the fact that the base language is regular, not necessarily CF. Every xy language on a regular base is in fact CF, from a small generalization of the result in Hayes (1965, Th. 4). Among the xy languages on strictly CF bases, however, there are infinitely many non-CF languages, as we shall show.

2. The formal and the empirical issues

Thus the class of xy languages, where x and y are distinct expressions drawn from a strictly CF language, does indeed have nontrivial members. We shall in fact show (in section 5) that this part of Chomsky's terse argument turns out to be soundly grounded mathematically (as PG). But some remarks need to be made about the data.

First, Chomsky's argument could still be formally flawed if the strings that he claims exemplify x and y in English are strings from a regular language. Additionally, his argument could be empirically unsound if English does not display the property he claims it does.

It seems clear that the set of strings corresponding to English noun phrases cannot be described in terms of a finite-state (FS) grammar. Precisely the same argument can be given for the non-FS nature of the set of English noun phrases have been given (in Chomsky 1957, for example) for the set of English sentences. So, for instance, two-part conjunctions like \textit{either...or...} and \textit{both...and...} must be nested in noun phrases just as they must be in sentences: \textit{either both a sculptor and a painter or an architect vs. *either a sculptor or a painter and an architect}. Center embedding of relative clauses also provides an argument for the strictly CF nature of language consisting of all English noun phrases: \textit{the dog that the rats that the cat hates see...} vs. \textit{*the dog that the rats that the cat hate sees...}. In such constructions, subjects and verbs must agree in a nested fashion and therefore display a mirror-image dependency that cannot be captured by any grammatical theory weaker than CF grammar.

From a formal point of view, then, there is nothing at all wrong with this type of language, and we shall consider the question of the empirical reality of the hypothesis that English is a strictly CF language.
Chomsky's thinking. However, we must agree with PG that the English comparative construction does not grammatically require nonidentity. In fact, we know of no natural language examples of required nonidentity that do not seem immediately explicable in terms of semantics or pragmatics, and we strongly suspect that there are none. For example, the frame

(3) It's not so much that \( S_1 \) as it is that \( S_2 \).

would seem to require nonidentity of \( S_1 \) and \( S_2 \), as indicated by the oddness of such examples as

(4) It's not so much that axolotls have gills as it is that axolotls have gills.

However, it is clear that this oddness is a function of the fact that when \( S_1 \) and \( S_2 \) are identical, the resulting sentence is empty and hence grammatically imperfect (by Grice's first maxim of quantity). Identity of \( S_1 \) has nothing to do with the deviance of such examples. Example (4) is exactly as deviant, and deviant in the same way, as (5),

(5) It's not so much that axolotls have gills as it is that they do.

If \( \text{they} \) is taken as referring to the class of axolotls; our (5) would be quite normal if \( \text{they} \) were taken to refer, say, to the class of lungfish.

But our suspicion might turn out to be wrong. If a language should turn up in which it is clear that there is grammatically required nonidentity between two expressions in a definable class of sentences, where these expressions themselves belong to a set that requires the power of CF grammar to specify, then that language, and hence natural language in general, might turn out to be provably beyond the power of CF phrase structure grammar. We return to this question in section 5.

### 3. Background to the Proof

We will present the argument for one strictly CF \( L_B \), and then observe that these results can be generalized.

This \( L_B \) is the set of all strings of the form \( a^n b^n \). An \( xy \) language \( L \) on this base is \( \{ a a^n b^n \beta a^m b^m \gamma; m \neq n \} \). If \( L \) is CF, so is \( L_1 = \{ c a b^e d a b^m f; m \neq n \} \). This follows from the fact that \( L \) is the homomorphic image of \( L_1 \) under the mapping in which \( c \) is replaced by \( \alpha \), \( d \) by \( \beta \), and \( f \) by \( \gamma \); the inverse homomorphic image of a CF language is itself CF (Hiscock and Ullman 1979, p. 132).
If $L_1$ is CF, so is $L_2 = \{a^nb^na^mb^m; m \neq n\}$. This follows from the fact that $L_2$ is the homomorphic image of $L_1$ under the mapping in which $c$, $d$, and $f$ are all replaced by the empty string $e$; the homomorphic image of a CF language is itself CF (Hopcroft and Ullman 1979, p. 132).

If $L_2$ is CF, so is $L_3 = \{a^nb^nc^md^m; m \neq n\}$. This follows from the fact that $L_2$ is the homomorphic image of $L_3$ under the mapping in which $c$ is replaced by $a$ and $d$ by $b$.

Finally, if $L_3$ is CF, so is $L_4 = \{a^nb^nc^m; m \neq n\}$. This follows from the fact that $L_4$ is the homomorphic image of $L_3$ under the mapping in which $d$ is replaced by $e$. We conclude that if $L$ is CF, so is $L_4$; and we will prove that $L_4$ is not CF (and so that $L$ isn’t either).

This proof, however, is not particularly easy to follow. Intuitively, it amounts to a demonstration that two interlocking conditions on the $a$’s, $b$’s, and $c$’s – that the length of the substring of $a$’s be the same as the length of the substring of $b$’s, and that the length of the substring of $b$’s be different from the length of the substring of $c$’s – cannot be satisfied simultaneously. The point might be clearer in terms of automata, using the well-known equivalence between the set of languages generated by CF grammars and the set of languages accepted by (nondeterministic) one-way pushdown-store automata.

Speaking very informally, such an automaton reads strings one symbol at a time from beginning to end and has access to a last-in-first-out stack. When the automaton starts on a string, $a^ib^ic^k$, it will need to store information about the length of the substring of $a$’s – this it could do by putting an $a$ on the stack for each of the $ia$’s – and then it will have to (i) check that the following substring of $b$’s has length $i$ AND ALSO (ii) store information about $i$ so that it can later determine that the substring of $b$’s is not the same length as the substring of $c$’s. But the automaton will use up the symbols on its stack in task (i) and have none left for task (ii). The obvious adjustment is to have the automaton put TWO $a$’s on its stack for each of the $a$’s in the string. Then it will not be able to manage task (i), since it has no way of knowing when it has worked its way down exactly half of the symbols in the stack. It cannot manage both tasks at once.

4. The proof for $L_4$

According to Ogden’s Lemma (Ogden, 1968; Hopcroft and Ullman 1979, pp. 129–30),² if $L_4$ is CF there is a constant $k$ such that if $z$ is any string in $L_4$, and we mark any $k$ or more positions of $z$ as ‘distinguished positions’
(6)(a) \( z \) can be decomposed as \( uvwxy \) in such a way that
(b) \( v \) and \( x \) together have at least one dp,
(c) \( uvx \) has at most \( kdp \)'s, and
(d) for all \( i \geq 0 \), \( uv^iwx^iy \) is in \( L_4 \).

Following the strategy of a sample proof given by Hopcroft and Ullman (1979, p. 130), we consider one particular string in \( L_4 \), namely

(7) \[ z = a^k b^k c^{2k+1+k} = uvwxy \]

and we let the \( k \) positions of the \( b \)'s be dp's.

We now propose to show that Ogden's Lemma requires that if \( z \) is in \( L_4 \), so is some string of the form \( a^kb^ce^l \), contrary to the defining conditions of \( L_4 \). Towards this end, we first show that in (6d) above \( v \) must consist entirely of \( a \)'s and \( x \) of \( b \)'s.

To begin with, note that \( v \) and \( x \) must each consist entirely of instances of a single symbol — for if they did not, then symbols would appear out of order in \( uv^iwx^iy \).

Next we observe that neither \( v \) nor \( x \) can be empty. If we suppose that \( v \) (alternatively, \( x \)) is empty, then by (6b) \( x \) (\( v \)) must have at least one dp, hence at least one \( b \), and so by our first observation must consist entirely of \( b \)'s. Then if \( v \) (\( x \)) is empty, all the \( a \)'s are in \( uv \) (\( u \)), and in \( uw^iwy \) \((uv^iwy)\) the number of \( a \)'s and \( b \)'s will not be identical for \( i \geq 2 \), as required by the original definition of \( L_4 \).

Next we see that \( v \) and \( x \) cannot both be strings of the same symbol. They cannot both be strings of \( a \)'s or \( c \)'s, for if they were, then \( uvw \) would contain no \( b \)'s, that is, no dp's, contrary to (6b). They cannot both be strings of \( b \)'s, for if they were, then \( u \) would have to contain all the occurrences of \( a \); but then the number of \( a \)'s and \( b \)'s will not be identical in \( uv^iwx^iy \) for \( i \geq 2 \).

It follows that \( v \) and \( x \) are strings of distinct symbols — either the symbols that must be unmatched in the strings of \( L_4 \) (that is, either \( a \) and \( c \), or \( b \) and \( c \)); or else the symbols that must be matched in the strings of \( L_4 \), that is, \( a \) and \( b \). But \( v \) and \( x \) cannot have unmatched symbols: If \( v \) consists entirely of \( a \)'s and \( x \) of \( c \)'s, then all the \( b \)'s are in \( w \); and if \( v \) consists entirely of \( b \)'s and \( x \) of \( c \)'s, then all the \( a \)'s are in \( u \). In either case, one of the matched symbols is in a fixed portion of \( uvwxy \) — in \( w \) or \( u \) — and the other is in the variable portion \( v \), and again the number of \( a \)'s and \( b \)'s will not be identical in \( uv^iwx^iy \) for \( i \geq 2 \).

The only remaining possibility is that \( v \) and \( x \) are strings of matched symbols, that is, that \( v \) consists entirely of \( a \)'s and that \( x \) consists entirely of \( c \)'s.
Now we consider the string $z' = uv^{2s+1}wx^{2s+1}y$. Since, by (6d), $uv'wx'y$ is not in $L_4$, since the number of $b$'s in $z'$ is identical to the number of $c$'s has $(k-p)$ b's, $wx^{2s+1}y$ has $(k-p) + (2k! + p) = (2k! + k)$ b's. But then $z'$ is not in $L_4$, since the number of $b$'s in $z'$ is identical to the number of $c$'s in $z'$.

It follows that $L_4$ is not CF, since its defining conditions and the requirements of Ogden's Lemma cannot be satisfied simultaneously. And so our original language $L$ is not CF either.

5. Generalization to (some) other $xy$ languages with strictly CF basis

The natural conjecture is that all $xy$ languages with strictly CF bases fail to be CF. Unfortunately, Ullian (1966) has shown that this is not so; the language $L' = \{a^i b^j d^k; \ i \neq j \text{ or } j \neq k\}$ is strictly CF, but Ullian shows that the $xy$ language $\{xcy; \ x, y \in L', \ x \neq y\}$ is CF.

Nevertheless, it is clear that there are infinitely many non-CF $xy$ languages. Any base language of the form $\{uv'wx'y; \ v, x \neq e, v \neq x\}$ is strictly CF, and an $xy$ language on this base can be shown not to be CF, by using a series of homomorphisms and inverse homomorphisms to argue that if the $xy$ language were CF, our original language $L$ would be too. Any base language of the form $\{uv'wu'y; \ v, w \neq e, w \neq v\}$ is strictly CF, and an $xy$ language on this base can be shown not to be CF by similar means.

The first of these facts is directly relevant to the issues in the first two sections of this note. Chomsky's (1963) discussion was incorrect in claiming that no $xy$ language with an infinite CF base is CF; all $xy$ languages with regular bases, and even some with strictly CF bases (like Ullian's $L'$), are CF. However, the $xy$ languages with the sorts of (strictly CF) bases Chomsky had in mind are probably not CF.

Against the background of our mathematical discussion, the argument that English is not CF because the set of comparative sentences of the type *John is better as NP$_1$ than he is as NP$_2$* is subject to a requirement that NP$_1$ and NP$_2$ be non-identical can be seen to depend in part on where the set of English NPs falls in the hierarchy of formal languages. If it is regular, then Chomsky's assertion is wrong. But of course he assumed that it was strictly CF, having provided arguments to this effect in Chomsky (1957), some of which we paraphrased in section 2 above. According to these empirical arguments, the set of English NPs is the sort of strictly CF base language that gives rise to non-CF $xy$ languages.

Consider, for example, the center embedding of relative clauses. To show that the set of NPs with center-embedded relatives is not regular, it is
sufficient to observe that the number of verbs in the relative clauses must
match the number of subject NPs in the relative clauses; all the NPs in the
set are of the form:

(7) \[ \text{NP [that NP]}'\text{V}. \]

But this is a base language of the form \{uv'wx'y; v, x \neq e, v \neq x\}, and we
noted above that an xy language on a base of this sort is indeed not CF.

We conclude that, in this case at least, a demonstration that there is a
grammatical requirement of nonidentity would in fact permit the con-
struction of an argument that English is not CF. The problem is not in the
mathematics, but in the linguistics.

Notes

* Zwicky's work on this note was supported in part by the System Development Foundation
through a grant to the Center for the Study of Language and Information, Stanford Univ.
Our thanks to Geoffrey Pullum for his comments on earlier versions of this note, and to
Stanley Peters for pointing us to relevant literature and saving us from grave error.

1 An anonymous reviewer of an earlier version of this note suggested that it might not be the
case that an xy language based on a non-CF language is itself non-CF. Certainly, as the
reviewer observed, an xx language based on a non-CF language need not be non-CF:
Assuming that the language consisting of strings of \(m^2 + n^2\) \(a\)'s is non-CF, the xx language
based on it, consisting of strings of \(m^2 + n^2 + p^2 + q^2\) \(a\)'s is certainly CF, indeed regular, since
it consists of all strings of \(a\)'s (given a theorem of number theory, due to Fermat, that all
natural numbers can be expressed as the sum of four squares).

2 We are indebted to two anonymous reviewers of an earlier version of this note for pointing
out that the pumping lemma for CF languages (Hopcroft and Ullman 1979, pp. 125–8) will
not serve to demonstrate that languages like \(L_4\) are not CF, and that something more
powerful (like Ogden's Lemma) is needed. Another language involving an inequali-
ty condition (which has also been claimed to be germane to the character of English as a CF
language, though Pullum (in press) argues that it is not) has been proved not to be CF by
Higginbotham (1984, Appendix), using Ogden's Lemma.

3 Languages of both these types have CF complements, but the complement of Ullman's
language is not CF. A conjecture worth pursuing is that an xy language with an infinite
deterministic strictly CF base is not CF.

References

Haines, L.: 1965, Generation and Recognition of Formal Languages, Ph.D. dissertation,
MIT.
Higginbotham, J.: 1984, 'English is Not a Context-Free Language', Linguistic Inquiry 15,
222–234.
Hopcroft, J. and J. D. Ullman: 1979, Introduction to Automata Theory, Languages and

Center for Advanced Study in the Behavioral Sciences
202 Junipero Serra Blvd
Stanford, CA 94305
U.S.A.