English and the Class of Context-Free Languages[1]

Paul M. Postal

IBM Thomas J. Watson Research Center
Post Office Box 218
Yorktown Heights, NY 10598

D. Terence Langendoen

Brooklyn College and the Graduate Center
City University of New York
33 West 42 Street
New York, NY 10036

0. Background

Let L range over all natural languages (NLs). For any L, one can consider two collections of strings of symbols, one consisting of all strings over the terminal vocabulary of L, call it W*(L), the other consisting of that always very proper subcollection of W*(L) consisting of all and only those members of W*(L) that are well-formed, that is, that correspond to sentences of L. Call the latter collection WF(L).

During the early development of generative grammar, a number of attempts were made to show, for various choices of L, that WF(L) was *not* a context-free (CF) string collection. These attempts all had, naturally, a common logical structure. First, it was claimed that there was some mathematical property P which, if possessed by some collection of strings, C, rendered C non-CF. Second, it was claimed that WF(C) had P, so the conclusion followed. Two sorts of criticisms can be, and have been, directed at such attempted demonstrations. One attacks the mathematical foundations and argues, for particular choices of P, that a collection manifesting P is not necessarily not CF. The other type of criticism admits that if a collection manifests a particular property P, it is thereby not CF, but contends that the WF(L)s claimed to manifest P in fact don't.

A survey of the various attempts, from roughly 1960 to 1982, to prove for various L that WF(L) is not CF is provided in Pullum and Gazdar (1982). These authors conclude, justifiably we believe, that for one or the other of the reasons mentioned above, none of these attempts, including those by the present authors, stand up. Despite widespread belief to the contrary, as of 1982 there had been no demonstration that there is some NL L for which WF(L) is not CF.[2]

However, Langendoen and Postal (1984) have obtained a result infinitely stronger than the claim that for some L, WF(L) is not CF. This work shows that for any L, WF(L) is a **proper class,** hence not a set, much less a recursively enumerable set. There is thus no question of WF(L) being CF. Moreover, WF(L) can then have no constructive characterization **(generative grammar)**, although there is no reason to doubt that it can be given a nonconstructive characterization. But the demonstration of Langendoen and Postal (1984) is based on principles that determine WF(L) includes nonfinite strings corresponding to nonfinite (transfinite) sentences. It is the existence of such sentences that places complete NLs beyond generative (constructive) characterization. Nevertheless, as noted in Langendoen and Postal (1984: 103), this novel result still leaves entirely open the question of whether that subpart of WF(L) consisting of all and only the well-formed *finite* strings in W*(L) is CF.

Let *F(inite)WF(L)* be that subcollection of WF(L) consisting of all and only the finite strings corresponding to the finite sentences of L. What follows shows that there are dialects of English, E1 and E2, such that:

[1] We thank J. Higginbotham for helpful comments on an earlier version of this paper.

[2] Recently, Higginbotham (1984) presents another argument that English is not CF. The formal part of the demonstration seems impeccable, but the factual premises are questionable; see Pullum (p. 182).

0362-613X/84/030177-05$03.00

1. Neither FWF(E1) nor FWF(E2) is CF.

The demonstration of (1) makes use of the following corollary of a theorem of Arbib (1969) about language intersections:

2. The Intersection Theorem
   Let L be a stringset and let R be a regular stringset.
   If L ∩ R is not a CF stringset, then neither is L.

One can then show that, for example, FWF(E1) and FWF(E2) are not CF by finding some regular set R such that R ∩ FWF(E1) and R ∩ FWF(E2) are not CF (Daly 1974).

1. The Sluicing Construction

The present demonstration that FWF(E), E ranging over various forms of English, is not CF is based on the **sluicing** construction, first discussed by Ross (1969) and more recently by van Riemsdijk (1978) and Levin (1982). Standard examples of this construction include:

3a. He stole something but it's not known what.
b. Someone stole the jewels and I can tell you who.
c. The police found him in some bar but the paper didn't say which one.

The sluicing construction has the following properties:

4a. It consists of n ($n \geq 2$) clauses, often, but not necessarily coordinate clauses joined by *but*
b. Each of the second to nth clauses counting from the left contains a *wh*-phrase, WH, which corresponds to a potential indefinite phrase found in the first clause.
c. WH has an interpretation equivalent to an entire *wh*-clause, WH+Z, and the 'missing parts' are understood as identical to the first clause minus the potential indefinite phrase.

The reason for the strange usage 'potential indefinite phrase' in (4b) is the existence of sluicing cases like (5b) and (6b) alongside (5a) and (6a):

5a. Martha was abducted by someone but we don't know by who(m).
b. Martha was abducted but we don't know by who(m).

6a. The doctors screamed for some reason but we don't know why.
b. The doctors screamed but we don't know why.

Depending on frameworks, one might analyze cases like (5b) and (6b) as involving invisible versions of indefinite phrases like those in the corresponding (5a) and (6a). But this matter need not concern us. We can concentrate on cases where alternatives like (5b) and (6b) are not possible, as in (7):

7a. Max visited someone but we don't know who.
b. *Max discussed Tom but we don't know who.
c. *Max discussed but we don't know who.

That is, we pick cases where WH can only be anaphorically connected to a *visible* element in the first clause. From this point on, references to the sluicing construction only denote cases of this restricted sort. The nature of the formal argument to be presented is such that this limitation in no way impugns the validity of the present result, since one can still construct a model of the intersection situation described at the end of the previous section.

In (3a), WH is *what,* the indefinite phrase corresponding to it is *something* and WH is understood as equivalent to the *wh*-clause *what he stole*, since *he stole* is the whole first clause minus the indefinite phrase. A similar analysis holds for (3b).

In (3), as in most of the examples in the literature previously illustrating the sluicing construction, the indefinite phrase in the first clause is an indefinite pronoun, and WH is a *wh*-pronoun. However, both phrases can consist of multi-word sequences:

8a. Sarah considered some proposals but it's unknown how many.
b. If any books are still left on the sale table, find out which ones.
c. The warehouse will ship us several typewriters but we have no idea how many machines.
d. A few physicians still use this drug and Sam can tell you how many doctors.
e. Joe discussed certain formulas but which formulas is uncertain.

Let us from this point on, for simplicity, limit attention to sluicing constructions consisting of only two clauses. Then it is possible to represent all the relevant cases schematically as follows:

9. V Q1 X1 W [Y Q2 X2 Z], where V, W, Y and Z are strings; Q1 is an indefinite quantifier or pronoun; X1 is the rest of the nominal quantified by Q1; Q2 is a *wh*-quantifier or pronoun anaphorically related to Q1; and X2 is the rest of the nominal quantified by Q2. Moreover, Q2 X2 (= WH) is understood as a *wh*-clause that contains material from V or W.

Table 1 presents the values of Q1, X1, Q2 and X2 in the examples in (8a-e). Henceforth, we further restrict the class of sluicing constructions under consideration, limiting attention only to examples like (8c-e), in which X1 and X2 *are neither empty nor pronouns.*

In (8c-e), the main **stress** on the *wh*-phrase can fall on either Q2 or on X2. If it falls on X2, then the *wh*-phrase can *not* be **anaphorically related** to the corresponding indefinite phrase in the first clause. While we cannot, and need not, give a theoretical account of "anaphorically related", informally it means that the potential reference of the *wh*-phrase is determined by that of its antecedent. Hence the lack of anaphoric connection in cases of stressed X2 means that in particular, in (8c), *machines* does not denote the same things that *typewriters* does in that sentence; in (8d), *physicians*

Table 1.

EXAMPLE	Q1	X1	Q2	X2
(8a)	some	proposals	how many	ϕ
(8b)	any	books	which	ones
(8c)	several	typewriters	how many	machines
(8d)	a few	physicians	how many	doctors
(8e)	certain	formulas	which	formulas

does not denote the same people that *doctors* does (the latter perhaps referring to nonmedical doctors); and in (8e) the two occurrences of *formulas* then denote different things (say, mathematical formulas in the first instance and baby milk formulas in the second). On the other hand, if phrasal stress falls on Q2, then the *wh*-phrase is anaphorically related to the corresponding indefinite phrase in the first clause. In (8c), *machines* is then taken to denote the same things that *typewriters* does; in (8d), *physicians* is then taken to denote the same people that *doctors* does and in (8e) the two occurrences of *formulas* denote the same objects, whether mathematical or nutritional.

Henceforth, we limit attention only to examples in which phrasal stress falls on Q2 (indicated by small caps). These are therefore structures where the *wh*-phrase in the second clause is anaphorically related to the corresponding indefinite phrase in the first clause.

It turns out that variants of English differ with respect to the class of *wh*-phrases that can be used anaphorically.[3] So, for many speakers, but not all, (8c-e) are fully acceptable with phrasal stress on Q2. For others, only (8e) is fully acceptable; in fact, (8c,d) are judged to be ungrammatical. For the latter dialect, henceforth referred to as E1, the subpart of the sluicing construction on which we have focused is subject to the constraint informally stated as in (10), henceforth referred to as the **strong matching condition** (SMC).[4]

10. If WH is the anaphoric *wh*-phrase in the second clause, then, if X2 is neither null nor a pronoun, the sequence of linguistic elements up to and including the *head noun* of X2 must be identical to the material up to and including the head noun of X1.

According to SMC, only the posthead modifiers of X1 and X2 can differ in E1, as in:

11. Joe discussed several attempts to grow corn on Mars but WHICH attempts is unknown.

If prehead modifiers differ, then the resulting structure is ungrammatical in E1:

12. E1*Joe discussed lots of curious proposals but everyone has forgotten WHICH proposals.

Now consider another dialect, call it E2, in which (8c-e) are fully grammatical when phrasal stress falls on Q2. For E2 speakers, the relation between X1 and X2 is governed by condition (13), which we refer to as the **weak matching condition** (WMC).

13. X2 is a possible anaphor of X1.

According to WMC, X2 can either be a complete repetition of X1, as in (8e); a synonym of X1, as in (8d); or a term whose denotation includes that of X1, as in (8c), (11) and (12), the last two examples being well-formed in E2. However, if X2 is not a possible anaphor of X1 and phrasal stress falls on Q2, then the resulting structure is ungrammatical even in E2. Since the following examples are ungrammatical in *both* E1 and E2, they are marked with double asterisks.

14a. **The warehouse will ship several machines to our office but we have no idea how MANY typewriters.
b. **A few physicians still use this drug and Sam can tell you how MANY nurses.
c. **Joe discussed certain formulas but WHICH equations is uncertain.

In (14a), *typewriters* cannot be used as an anaphor for *machines*, presumably because the reference of the former fails to subsume that of the latter. This judgement is rendered by E2 speakers even for contexts in which the words *typewriters* and *machines* are otherwise used interchangeably, such as an office with limited word processing equipment. This shows that one is dealing here with a grammatical restriction, not a pragmatic property that varies with context. Similarly, in (14b), *nurses* cannot be used as an anaphor for *physicians*, since some physicians are not nurses and some nurses are not physicians; again this is true even in a context in which all the nurses under discussion happen to be physicians and vice versa. Finally, *equations* cannot be used as an anaphor for *formulas* in (14c), since again an equation is only a certain kind of formula (statements of inequality are also formulas). One who judges that (14c) is in fact grammatical might well do so under the mistaken belief that all mathematical formulas are equations. Alternatively, one could assume that such a judge has a different dialect of English, with different anaphoric conditions.

Now consider examples of the sluicing construction in which a **compound noun** occurs as X1. If X2 exactly matches X1, then the result is grammatical in both E1 and E2. If only the head of the compound occurs as X2, then the results are always ungrammatical in E1, and either

[3]We have not investigated whether this correlates with more general differences in anaphoric usages for these distinct forms of the language. Moreover, this is not relevant to the present demonstration.

[4]E1 is the dialect of the first author.

grammatical or ungrammatical in E2, depending on the relation between X1 and X2. As before, double asterisks mark examples that are ungrammatical in both E1 and E2.

15a. Joe discussed some candy store but it's not known WHICH candy store.
b. E1*Joe discussed some candy store but it's not known WHICH store.
c. **Joe discussed some fire escape but it's not known WHICH escape.
d. **Joe discussed some bourbon hater but it's not known WHICH hater.
e. **Joe discussed some bourbon lover but it's not known WHICH lover.

The whole compound can be used as an anaphor for itself in both E1 and E2, as in (15a). But *store* cannot be used as an anaphor for *candy store* in (15b) in E1, since SMC is not satisfied. On the other hand, *store* can be used as an anaphor for *candy store* in (15b) in E2, presumably since a candy store is a certain kind of store; that is, *candy store* is an endocentric compound. However, *escape* cannot be used as an anaphor for *fire escape*, even in E2, since a fire escape, which is a certain kind of physical object, is not an escape, which is a certain kind of event; that is, *fire escape* is an exocentric compound. Finally, *hater* and *lover* cannot be used as anaphors for *bourbon hater* or *bourbon lover* in (15d,e), since the agentive noun *hater* is used only in compounds and lover; by itself has a limited (sexual) meaning which makes it unsuitable as an anaphor for compounds such as *bourbon lover*

An important consequence for the present discussion is that if one limits the vocabulary over which sluicing constructions are formed in the right way, the conditions of linkage in E1 and E2, though intensionally distinct, become extensionally identical. That is, for fixed cases, SMC and WMC have the same consequences. For any such situation, one can thus equate them and refer simply to the **matching condition** (MC).

It is possible to embed English compounds within compounds; in particular, there are well-formed compounds such as *bourbon hater lover* 'one who loves bourbon haters', *bourbon lover hater* 'one who hates bourbon lovers', *bourbon hater lover hater* 'one who hates those who love bourbon haters', etc. Now, if any such compound occurs as X1 in a sluicing construction, then for speakers of both E1 and E2 the only possible anaphor *drawn exclusively from the vocabulary used to construct the compound* that can occur as X2 is the whole compound itself.

16a. Joe discussed some bourbon hater lover but it's not known WHICH bourbon hater lover.
b. **Joe discussed some bourbon hater lover but it's not known WHICH hater lover.
c. **Joe discussed some bourbon hater lover but it's not known WHICH lover.

It follows that if attention is limited to instances of the sluicing construction like (16), in which the only possible anaphoric *wh*-expressions are whole compounds, then SMC and WMC are equivalent, permitting one to speak simply of MC with no loss of accuracy. Given this fact, we are now in a position to demonstrate simultaneously that neither FWF(E1) nor FWF(E2) is CF.

2. The Proof

We first define the following notion:

17. A **copying language** is any language of the form:

L {cxdxe | x ϵ (a,b)* and a,b,c,d,e are fixed strings}

Given that English contains a sluicing construction characterized as in section 1, one can prove that FWF(E), E ranging over E1 and E2, is not CF by means of the Intersection Theorem of (2) and the fact that copying languages are not CF (Langendoen 1977). To prove that FWF(E) is not CF, one must find a regular language R whose intersection I with FWF(E) is not CF. Such an R is given in (18):

18. R = {Joe discussed some bourbon x but WHICH bourbon y is unknown | x,y ϵ (hater, lover)*}

Since R is a concatenation of regular languages, it is itself regular.

Now consider the intersection of R with FWF(E). The matching condition on sluicing constructions guarantees that this intersection is the copying language I in (19):

19. I = {Joe discussed some bourbon x but WHICH bourbon x is unknown | x,y ϵ (hater, lover)*}

Since I is not CF, by the Intersection Theorem, neither is FWF(E).

As has been stressed, limitations on the vocabulary render the distinct sluicing conditions of E1 and E2 equivalent over certain subcollections of sluicing cases. It follows that the demonstration just presented holds not only for E1 and E2 but more generally for any variant of English whose matching condition for sluicing, even if different from that of both E1 and E2 for the full collection of English sentences, has the same extension for the language R of (18). We see no current reason to doubt that this will include *every* variant of English.

3. Conclusion

Gazdar (1983: 86), summarizing inter alia the conclusions of Pullum and Gazdar (1982), makes the following claims:

20a. "There is no reason, at the present time, to think that NLs are not CFLs."
b. "There are good reasons for thinking that the notations we need to capture significant syntactic generalisations will characterise CF-PSGs, or some minor generalisations of them, such as Indexed Grammars."

But as Langendoen and Postal (1984) shows, NLs are proper classes not sets, so the question of WF(L)s as wholes being CF no longer arises. Restricting attention to FWF(L)s, the result of Section 2 shows that for any dialect E of English for which MC holds, FWF(E) is not CF.

Since, however, neither FWF(E) nor any other FWF(L) has been shown to lie outside the domain of indexed languages (ILs) in the sense of Aho (1968), it would appear that one can conclude that while the collection of all sentences in an NL K is a proper class, FWF(K) is an IL. Consequently, a correct account of NL grammars must be such that a proper grammar for K specifies K as an appropriate proper class[5] and entails that FWF(K) is an IL. A grammatical theory with just these properties remains to be constructed.

[5]By 'appropriate', we mean one which satisfies inter alia the axiom called Closure Under Coordinate Compounding of Sentences in Langendoen and Postal (1984: 53). This is necessary for the proof that NLs are proper classes.

References

Aho, A.V. 1968 Indexed Grammars–An Extension of Context-Free Grammars, *Journal of the Association for Computing Machinery* 15: 647-671.

Arbib, M. 1969 *Theories of Abstract Automata.* Prentice-Hall, Englewood Cliffs, New Jersey.

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

Gazdar, G. 1983 NLs, CFLs and CF-PSGs. In: Sparck Jones, K. and Wilks, Y., Eds., *Automatic Natural Language Parsing.* Ellis Horwood Ltd., West Sussex, England.

Higginbotham, J. 1984 English is Not a Context-Free Language. *Linguistic Inquiry* 15: 119-126.

Langendoen, D.T. 1977 On the Inadequacy of Type-2 and Type-3 Grammars for Human Languages. In: Hopper, P.J., Ed., *Studies in Descriptive and Historical Lingistics.* John Benjamins, Amsterdam, Holland.

Langendoen, D.T. and Postal, P.M. 1984 *The Vastness of Natural Languages.* Basil Blackwell, Oxford, England.

Levin, L. 1982 Sluicing: A Lexical Interpretation Procedure. In: Bresnan, J., Ed., *The Mental Representation of Grammatical Relations.* The MIT Press, Cambridge, Massachusetts.

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

Ross, J. R. 1969 Guess Who. In: Binnick, R. et al., Eds., *Papers from the Fifth Regional Meeting Chicago Linguistic Society.* University of Chicago, Chicago, Illinois.

van Riemsdyk, H. 1978 *A Case Study in Syntactic Markedness: The Binding Nature of Prepositional Phrases.* Foris Publications, Dordrecht, Holland.