Simultaneous-Distributive Coordination and Context-Freeness

Michael B. Kac

Department of Linguistics
University of Minnesota
Minneapolis, Minnesota 55455

Alexis Manaster-Ramer

Department of Computer Science
Wayne State University
Detroit, MI 48202

William C. Rounds

Department of Electrical Engineering and Computer Science
University of Michigan
Ann Arbor, MI 48109

English is shown to be trans-context-free on the basis of coordinations of the *respectively* type that involve strictly syntactic cross-serial agreement. The agreement in question involves number in nouns and reflexive pronouns and is syntactic rather than semantic in nature because grammatical number in English, like grammatical gender in languages such as French, is partly arbitrary. The formal proof, which makes crucial use of the Interchange Lemma of Ogden et al., is so constructed as to be valid even if English is presumed to contain grammatical sentences in which *respectively* operates across a pair of coordinate phrases one of whose members has fewer conjuncts than the other; it thus goes through whatever the facts may be regarding constructions with unequal numbers of conjuncts in the scope of *respectively,* whereas other arguments have foundered on this problem.

> ***respective(ly).*** **Delight in these words is a widespread but depraved taste.**
> **Fowler (1937: 500)**

Introduction

Pullum and Gazdar (1982) systematically review and critique a large number of arguments for trans-context-freeness of natural languages,[1] finding each one defective conceptually, empirically, or mathematically. Among these are various ones (e.g., that of Bar-Hillel and Shamir (1960)[2]) appealing to the existence in English of sentences like

(1) John and Bill dated Mary and Alice respectively.

Pullum (1984) cites a number of more recent arguments, involving languages other than English, which do appear to establish their trans-context-freeness and remarks (p. 117), in connection with a suggestion regarding Swedish gender agreement, that if this type of agreement can be shown to be a purely syntactic matter, then sentences analogous to English instances of the schema *The* **N, N,** ... *and* **N** *are respectively* **A, A** ... *and* **A′**, might provide the basis of an argument for trans-context-freeness of the language (or of any other with similar facts). In this paper, we shall produce a rigorous argument along comparable lines to show that English is trans-context-free (trans-CF), though we shall rely on facts regarding grammatical number rather than gender.

The relevance of a strictly negative result such as the one we have obtained is not restricted to the narrow question of where natural languages do (or don't) place in the Chomsky hierarchy. Given that proving the trans-

0362-613X/87/010025-30$03.00

context-freeness of particular natural languages has turned out to be considerably more difficult than anyone had expected it to be, and that solutions to difficult problems are likely to bear fruit outside the parochial confines of the original problem area and to call attention to hitherto unnoticed facts, an exercise such as this goes well beyond lily gilding or dead horse beating. To develop our argument, for example, we shall turn the spotlight on the linguistics phenomenon of arbitrary number, something which is rarely mentioned in standard treatment of grammatical phenomena (in vivid contrast to arbitrary gender), but which turns out to be more than a mere curiosity.

The mathematical approach that we employ is also noteworthy in making crucial use of the Interchange Lemma for CFLs (Ogden et al. 1985) and of a "separation" technique that allows trans-context-freeness to be demonstrated by showing that certain strings are included in a language while others are excluded. Neither of these has, to our knowledge, been used in a natural language context before. The interest of the separation method in particular lies in the way in which it simplifies the following problem. Since natural languages are large, complex, and (most important) lacking in antecedent definitions, the only practical way to argue about their mathematical properties is to examine sublanguages. If one is not careful, however, one runs the risk of committing the "trickle-up" fallacy, which consists in showing that a certain set S has a property P and then attributing P to some proper superset of S. The usual way of circumventing this difficulty is to capitalize on closure properties of languages under intersection with a regular set; the separation technique provides an alternative in cases where appeal to such closure properties is not sufficient (or at least not obviously so).

Partly in the interest of a terminology free of English bias, we call constructions like (1) **simultaneous-distributive (SD) coordinations.** This label reflects the fact that a sentence such as (1) can be "unpacked" to yield, *salva veritate,* a coordination of noncoordinate sentences by means of the following procedure:

- first, put a copy of the verb directly before each NP in the second coordinate phrase that is not already immediately preceded by a verb, thus creating a coordination of VPs;
- then "distribute" the NPs in the first coordinate phrase among the VPs by simultaneously associating a copy of each of the former with exactly one of the latter, namely the one in the corresponding positions;
- finally, suppress the first coordinate phrase, the first *and* and *respectively.*

1 Mathematical Techniques for Establishing Trans-Context-Freeness

We shall rely here on a number of established mathematical results which, taken together, give us a way of establishing trans-context-freeness for a language with certain syntactic properties.

Theorem 1 (Bar-Hillel et al. 1961)

The set of context-free languages is closed under homomorphism.

Theorem 2 (Interchange Lemma, Ogden et al. 1985)

Let L be a CFL, and let L_n be the set of length n strings in L. Then there is a constant C_L such that for any n, any nonempty subset Q_n of L_n, and any integer m such that $n \geq m \geq 2$, the following holds: Let $k = \lceil \| Q_n \| / (C_L n^2) \rceil$, where $\lceil x \rceil$ denotes x rounded up to the nearest integer, and $\| Q_n \|$ is the cardinality of Q_n. Then there are k distinct strings $z_1, \ldots, z_k$ in Q_n such that z_i can be written $w_i \ x_i y_i$ for $1 \leq i \leq k$, and:

(i) $| w_i | = | w_j |$ for all $i, j \leq k$;

(ii) $| y_i | = | y_j |$ for all $i, j \leq k$;

(iii) $m \geq | x_i | > m/2$;

(iv) $| x_i | = | x_j |$ for all $i, j \leq k$; and

(v) $w_i \ x_j \ y_i \in$ L for all $i, j, \leq k$.

Since this result is likely to be unfamiliar to some readers, we shall provide some commentary that should prove helpful in following the remainder of the presentation.

Less forbiddingly stated, the Interchange Lemma says (in part) that in a CFL it is possible, for any n, to find at least two strings of length n with internal parts of the same length that can be exchanged for each other to produce strings that are also in L, providing that there are at least two distinct strings of length n. This makes it possible to prove the trans-context-freeness of a certain kind of language (what kind will be stated in a moment) by showing that once n become sufficiently large, the possibility of interchange no longer exists. For this strategy to work, it is required that the cardinality of L_n grow very rapidly as a function of n, which we can illustrate with the case of the copying language over the vocabulary $\{a, b\}$ (call this language CP): For every $n \leq 2$, $\| CP_n \| = 2_{n2}$, thus growing exponentially, while $C_{CP} n^2$ grows only polynomially; the necessary conditions for use of the Interchange Lemma to prove trans-context-freeness are thus satisfied by this case.

Making use of the Interchange Lemma, we have proved the following further result:

Theorem 3.

Let H = $\{xy \mid x \in \{a, b\}^*, y \in \{c, d\}^*, y = h(x)$, where $h(a) = c, h(b) = d\}$ and G = $\{xy \mid x \in \{a,b\}^*, y \in \{c, d\}^*, |x| \neq |y|\}$ Then any set I is trans-CF if I is a superset of H and a subset of G ∪ H.

The proof is presented as an appendix to the paper. As an immediate corrollary, we obtain the following result:

Theorem 4.

Let G, H, and I be as defined in Theorem 3, and let $J = \{xy \mid x \in \{a, b\}^*, y \in \{c, d\}^*, |x| = |y|\}$ and $K = J - H$. Then any set L containing H and disjoint from K is trans-CF.

Proof: Intersect L with the regular set $H = \{xy \mid x \in \{a, b\}^*, y \in \{c, d\}^*\}$ and let $N = L \cap M$. H is a subset of N, K is disjoint from N, and the only other strings that might be in N are those in G. Therefore, N contains H and some subset (possibly empty) of G, and is trans-CF by Theorem 3. Since the intersection of any CFL with any regular set is CF, L is trans-CF.■

Theorem 4 says that an arbitrary language L is trans-CF if it meets the following conditions:

- It includes H.
- It excludes every string not in H that is nonetheless divisible into two equal parts, the first over $\{a, b\}$ and the second over $\{c, d\}$. (This is the set K.)

In order to apply these results, we sill actually need to consider not quite the sets G through N as defined above but the corresponding sets G' through N', where the latter differ from the former in including only strings whose x and y parts are of at least length 2. The subtraction of a finite subset obviously changes nothing essential, and Theorems 3 and 4 will hold, mutatis mutandis, of sets G' through N'. Hence:

Theorem 5.

Let H' and K' be as defined above. Then any set L' containing H' and disjoint from K' is trans-CF.

Our strategy in applying these results to English will be to show that there is a subset F of English that can be homomorphically mapped to some L', and that F is the intersection of English with a regular language. This suffices to show that English itself is trans-CF.

2 Grammatical Number Agreement in English

Our empirical argument rests on the claim that number agreement between reflexive pronouns and their antecedents is a syntactic phenomenon in English. For example, the string

(2) *The girl likes themselves.

must be considered ungrammatical rather than merely semantically ill-formed by virtue of the impossibility (because of number incompatibility) of supplying an intraclausal antecedent for the reflexive pronoun. The reason that this is so has to do with a fact about grammatical number in English that has not been generally recognized; namely, that it is, like grammatical gender in languages such as French and German, partly arbitrary. This can be shown by a number of different kinds of examples, among them the following. First, there are synonym pairs in English, each consisting of a grammatically singular member and a grammatically plural one. A partial list is given in the table below.[3]

SINGULAR	PLURAL
apparel	clothes
forest	woods
underwear	underpants
car	wheels
kibble	crunchies
location	whereabouts
merchandise	goods
Pamir	Pamirs*
Hellespont	Dardanelles
corpse	remains
pant**	pants
hosiery	stockings
military	armed forces
issue	offspring

* a mountain range in Central Asia

** as used in the garment trade

Further, these examples can be elaborated in various ways. For example, names of some mountain ranges are strictly singular *(Caucasus, Hindu Kush)*, while those of others are strictly plural *(Alps, Rockies)*; items from similar semantic fields may vary as to their grammatical number properties (compare *odds-probability, wheat-oats, yoghurt-curds, pasta-noodles, mush-grits* (in some dialects), *Granola-Rice Krispies*). Note further that there is dialect variation regarding the grammatical number of certain collective nouns (such as *government* and *company*), which are strictly singular in American English, but which can be used as plurals in British English.

A further phenomenon on which we shall capitalize is the existence in English of an idiomatic way of expressing the ease with which an activity can be performed involving the use of reflexive constructions, as illustrated by, for example, *This land will rent itself,* or *These woods will sell themselves.* With this in mind, compare now

(3) This land and these woods can be expected to rent itself and sell themselves respectively.

(4) *This land and these woods can be expected to rent themselves and sell itself respectively.

It is clear that strings like (3), in which each reflexive pronoun agrees with the corresponding noun, are grammatical, while those in (4), in which each pronoun disagrees with the corresponding noun, are not. This fact will be the basis of our demonstration that English is trans-CF.

3 Argument Regained

Let A = {*{this land, these woods} and {this land, these woods}*$^+$ *can be expected to {{rent, sell} {itself, themselves}}*$^+$ *and {{rent, sell} {itself, themselves}} respectively*}, and note that A is regular. Now let B be the subset of A that satisfies the following condition:

In case the number of occurrences of members of {*this land, these woods*} is equal to the number of occurrences of members of {*itself, themselves*}, then for all $i \geq 1$, if the *i*th noun in the string is *land,* then the *i*th pronoun is *itself* and if the *i*th noun is *woods,* then the *i*th pronoun is *themselves* (i.e., number agreement obtains between the nouns and the reflexive pronouns).

Now let C = A − B. Every string in C contains exactly as many pronouns as it does nouns, but for some $i \geq 1$, the *i*th pronoun fails to agree in number with the *i*th noun. Finally, let D be the subset of B consisting of just those strings that contain exactly as many nouns as pronouns.

It is clear that D is part of English, and that C is disjoint from English, inasmuch as D exhibits the required number agreement and C does not. If D were the intersection of English with the regular set A, then the result that English is trans-CF would follow immediately.[4] However, things are not that simple, and it is conceivable that the intersection of English with A is some proper superset F of D that is a subset of B (possibly B itself). The point of uncertainty here is the status of strings of A with unequal numbers of occurrences of nouns and pronouns, such as the following:

(5) This land, these woods, and this land can be expected to rent itself and sell themselves respectively.

(6) This land and these woods can be expected to rent itself, sell themselves, and rent itself respectively.

While it might seem that such strings are ungrammatical, this assumption is called into question by Pullum and Gazdar's (1982) observation that there is no syntactic constraint in English governing the number of conjuncts in SD-coordinations. Thus, contrary to conventional wisdom, there are perfectly well-formed SD-co-ordinations with mismatched numbers of conjuncts; for example, *The last two people in this picture live in Columbus and Chicago respectively.* This undermines a number of older arguments that English is trans-CF that crucially assume that grammatical SD-coordinations must have equal numbers of conjuncts. In light of this, it may be that strings like (5-6) are to be considered syntactically well-formed, albeit lacking sensible interpretations, and so an argument presupposing the contrary cannot be used to show that English is trans-CF.

We now show that Theorems 3, 4, and 5 allow us to get around this obstacle by, in effect, ignoring the strings with mismatched numbers of conjuncts in constructing our argument. If strings like (5-6) are grammatical, then the intersection of English with the regular language A is not the trans-CF language D but some proper superset F thereof. Theorem 5 tells us in effect that, no matter what its exact identity, if there is a sublanguage of English homomorphic to H′ but none homomorphic to K′, then English is trans-CF. This "separation" strategy yields the conclusion that so long as English contains D and excludes C, it is trans-CF.

Recall now that, according to our definition, B includes strings like (5-6), along with sentences like (3) but, crucially, excludes (4) and all strings like it. To be precise, B includes all strings of A which, like (3) have exactly as many nouns as pronouns *and* cross-serial number agreement, but also all strings in A that, like (5) and (6), have more pronouns than nouns or vice versa. In virtue of Theorem 5, we will be able to show that English is trans-CF no matter what position we take on the grammaticality of strings like (5-6), so long as there are no English sentences in the subset C of A consisting of strings in which there are as many nouns as pronouns but at least one of the pronouns fails to agree with the corresponding noun in the first part of the string. Thus, the intersection of English with A is some subset F of B that is disjoint from C; it is of no consequence whether F is equal to all of B, or only to D, or to some proper subset of B that is a proper superset of D.

We now define the homomorphism *h* such that

h (this land) = *a*

h (these woods) = *b*

h (itself) = *c*

h (themselves) = *d*

For all *z* ∈ {*can be expected to, sell, rent, and, respectively*}, *h(z)* → ϕ

This homomorphism maps F to L′. Since the CFLs are closed under homomorphism, and L′ is trans-CF, F is trans-CF. And since the CFLs are also closed under intersection with regular sets, and the intersection of English with the regular set A has turned out to be trans-CF, it follows that English is also trans-CF.

It should be immediately apparent that a similar strategy can be applied in instances such as the one mentioned in Section 1, where grammatical gender agreement is involved, providing that the language in question has instances of arbitrary gender. So, for example, we can construct for French a sublanguage parallel to D consisting of sentences like

(7) Cette nation et ce pays sont respectivement une alliée et un associé des Etats-Unis.

'This nation and this country are respectively an ally and a partner of the United States.'

In this example, we capitalize on the fact that the inanimate nouns *nation* 'nation' and *pays* 'country' belong to different gender classes, reflected in the predicate nominals *une alliée* 'an ally' and *un associé* 'a partner'.[5] Inversion of the predicate nominals yields an ungrammatical string, but corresponding inversion of the subjects restores grammaticality:

(8) *Cette nation et ce pays sont respectivement un associé et un alliée des Etats-Unis.

(9) Ce pays et cette nation sont respectivement un associé et une alliée des Etats-Unis.

Comparable examples can be constructed in other languages, a case in point being the Polish sentence

(10) Francja i Kongo są przeciwniczką względnie zwolennikiem traktatu.

which translates literally as 'France and (the) Congo are opponent respectively supporter (of the) treaty'; here, as in the French example, the two nouns in the subject phrase are of different grammatical genders, and are matched with corresponding gender-compatible nouns in the predicate phrase. Example (11) is ungrammatical (notice the suffixes of the predicate nouns) while (12) is grammatical again.

(11) *Francja i Kongo są zwolennikiem względnie przeciwniczką traktatu.

(12) Kongo i Francja są zwolennikiem względnie przeciwniczką traktatu.

4 Conclusion

We would like to close by pointing out the importance and interest of,the following question: Given that it is logically possible for a language to have an operator just like English *respectively* except that the conjuncts to be linked with each other are paired in center-embedding rather than cross-serial fashion, and given that the properties of such an operator can be characterized by formal apparatus apparently more elementary than what is required to characterize *respectively,* why do operators of this seemingly more elementary type appear not to exist in any natural language? The use of *apparently* is important here: from the standpoint of the Chomsky hierarchy, nesting is less complex than mutual intercalation, in the sense that the type of grammar required to handle the former is more restricted than the type required to handle the latter, but the possibility is always open that this type of complexity is not germane to human psychological capacity. We strongly suspect that this is the case (see Manaster-Ramer and Kac 1985), though that is a topic for another time.

Appendix: Proof of Theorem 3

Assume that $I \subseteq G \cup H$ is context-free and apply the Interchange Lemma to

$$Q_{2n} = \{xh(x) \mid x \in \{a, b\}^n\}$$

Since $H \subseteq I$, $Q_{2n} \subseteq I_{2n}$, the set of length $2n$ substrings in I. Choose n suitably large (for the exact choice, see the proof of Claim 1 below), and let $m = n$. Let $k = \lceil \|Q_{2n}\| / (C_I(2n)^2) \rceil$ be the number defined in the lemma. We get k distinct z_i in Q_{2n} satisfying the conclusion of the lemma. Our result will follow from the next two claims:

> **Claim 1.** Let $x_i, \ldots, x_k$ be the middle parts of $z_i, \ldots, z_k$ respectively. Then there are i and j such that $x_i \neq x_j$, provided that $_n$ is suitably chosen.
>
> **Claim 2.** If $x_i \neq x_j$, then $w_i x_j y_i$ is not in $G \cup H$ and, a fortiori, not in I.

Proof of Claim 1. Suppose that all the x_i were equal. By (iii) of the Interchange Lemma, $|x_i| > n/2$. Therefore, at least $n/4$ characters from the x_i are in the $\{a, b\}$ half or in the $\{c, d\}$ half of z_i. We may assume the former possibility since the argument works exactly the same way in the latter case. Each z_i is determined by the n characters of its $\{a, b\}$ half and, by supposition, at least $n/4$ of these characters are fixed. So there can be at most $2^{n-n/4} = 2^{3n/4}$ strings $z_1, \ldots, z_k$. But $\|Q_{2n}\| = 2^n$, and if n is chosen so that $2^{3n/4} < 2^n/(C_I n^2) = \|Q_{2n}\|/(C_I n^2)$, we contradict the Interchange Lemma, which says that all the z's are distinct. Note that n can always be chosen this way, no matter what C_I is, by elementary inequalities from college algebra and calculus.

Proof of Claim 2. Observe first that interchanging x_i with x_j does not produce a string in G. The substrings x_i and x_j disagree in some position, which is also a position in one half or the other of the strings z_i and z_j. Thus the matching position in z_i and z_j (in the other half of the word) does not occur in x_i or x_j because $|x_i| = |x_j| \leq n$, by (iii) of the Interchange Lemma, so interchanging x_i and x_j produces a string not in H.

Since Claim 1 and Claim 2 violate clause (v) of the Interchange Lemma, I cannot be context-free.■

References

Bar-Hillel, Y. and Shamir, E. 1960 Finite-State Languages: Formal Representations and Adequacy Problems. In Bar-Hillel, Y., Ed., *Language and Information.* Addison Wesley, Reading, Massachusetts: 87-98.

Bar-Hillel, Y.; Perles, M.; and Shamir, E. 1961 On Formal Properties of Simple Phrase Structure Grammars. *Z. Phonetik. Sprachwiss. Kommunikationsforsch.* 14: 143-172.

Chomsky, N. 1963 Formal Properties of Grammars. In Luce, R.D.; Bush, R.R.; and Galanter, E., Eds., *Handbook of Mathematical Psychology, Volume II.* John Wiley and Sons, New York, New York: 323-418.

Fowler, W. 1937 *A Dictionary of Modern English Usage.* Clarendon Press, Oxford.

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

Manaster-Ramer, A. and Kac, M.B. 1985 Formal Models and Linguistic Universals. Paper read at the Symposium on Language Typology and Universals, University of Wisconsin at Milwaukee.

Ogden, W.; Ross, R.J.; and Winklmann, K. 1985 An "Interchange Lemma" for Context-Free Languages. *SIAM Journal of Computing* 14: 410-415.

Postal, P. 1964 Limitations of Phrase Structure Grammars. In Fodor, J.A. and Katz, J.J., Eds., *The Structure of Language.* Prentice-Hall, Englewood Cliffs, New Jersey: 137-154.

Pullum, G.K. 1984 Syntactic and Semantic Parsability. *Proceedings of the 10th International Conference on Computational Linguistics:* 112-122.

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

Notes

[1] We use the term **trans-context-free** in preference to **non-context-free**, to distinguish the class of languages outside the weak generative capacity of type 2 grammars.

2. This argument is reiterated in Postal (1964); however the relevance of SD-coordination to the question of context-freeness was evidently first noticed in 1959 by Ray Solomonoff in personal communication to Chomsky (see Chomsky 1963).
3. Some of these examples were suggested to us by Rosemarie Whitney.
4. See, for example, the argument, due to H. Brandt Corstius, cited in Levelt (1974: 31-32).
5. These nouns were selected because they have counterparts of the opposite gender: thus, corresponding to *alliée* we have *allié,* and corresponding to *associé* we have *associée.* The masculine and feminine forms of 'partner' and 'ally' are phonologically indistinguishable in isolation, but can be distinguished in the context of preceding indefinite articles, as our examples show.