SRI International

EVIDENCE AGAINST THE CONTEXT-FREENESS OF NATURAL LANGUAGE

Technical Note 330

June 1984

By: Stuart M. Shieber, Computer Scientist
Artificial Intelligence Center
Computer Science and Technology Division
and
Center for the Study of Language and Information
Stanford University

To appear in *Linguistics and Philosophy.*

This research has been made possible in part by
National Science Foundation Grant No. IST 83-07893
and by a gift from the System Development Foundation,
and was also supported by the Defense Advanced
Research Projects Agency under Contract N00039-80-C-0575
with the Naval Electronic Systems Commands.

SRI International®

333 Ravenswood Ave. • Menlo Park, CA 94025
(415) 326-6200 • TWX: 910-373-2046 • Telex: 334-486

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
JUN 1984		**00-06-1984 to 00-06-1984**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
Evidence Against the Context-Freeness of Natural Language	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **SRI International,333 Ravenswood Avenue,Menlo Park,CA,94025**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT **unclassified**	b. ABSTRACT **unclassified**	c. THIS PAGE **unclassified**		**14**	

1. Introduction[1]

In searching for universal constraints on the class of natural languages, linguists have investigated a number of formal properties, including that of context-freeness. Soon after Chomsky's categorization of languages into his well-known hierarchy [Chomsky, 1963], the common conception of the context-free class of languages as a tool for describing natural languages was that it was too restrictive a class—interpreted *strongly* (as a way of characterizing structure sets) and even *weakly* (as a way of characterizing string sets).

The issue was brought back to the attention of linguists a few years ago, however, by Gerald Gazdar's arguments for a context-free phrase-structure theory of syntax [Gazdar, 1982]. Subsequently, Gazdar and Geoffrey K. Pullum [1982] chronicled common thinking on the issue, and argued compellingly against all previous published arguments maintaining the weak non-context-freeness of natural language. Since then, to the author's knowledge, no proof of the weak non-context-freeness of natural language has been forthcoming.[2]

However, one of the arguments discussed by Gazdar and Pullum—that concerning the Dutch cross-serial clause construction [Bresnan, *et al.*]—came quite close. The class of structures propounded on linguistic grounds for grammatical subordinate clauses with the cross-serial construction was demonstrated to be non-context-free. That is, although the string set of Dutch was

[1] The author would like to thank Beat Buchmann, Mark Domenig, Hans Huonker and Patrick Shann for their patience in providing the Swiss-German data, and the researchers at the Dalle Molle Institut pour les Etudes Semantiques et Cognitives for providing the impetus and opportunity to pursue this study. Special thanks go to Thomas Wasow for his extensive and continued support of this research.

[2] Several new arguments have been proposed recently. Those of Higginbotham [1984] and of Langendoen and Postal have been convincingly refuted by Pullum. An argument for the non-context-freeness of the lexicon has been proposed by Culy. Its status is unclear at this time.

not (and could not be) shown to be ungenerable by a context-free grammar, the constituent structure set nevertheless was—if Bresnan *et al.* are right about the linguistic motivation for those structures. Of course, their demonstration relied greatly upon linguistic arguments as well as formal language theory and, in fact, several authors have presented alternative analyses [Culy, 1983,Joshi, 1983,Thompson, 1983]. Although all linguistically motivated analyses have been strongly non-context-free, one in particular [Culy, 1983] was weakly context-free.[3]

This paper offers evidence for the *weak* non-context-freeness of natural language. Using data collected from native Swiss-German speakers, we will provide a formal proof of the weak non-context-freeness of Swiss German. In doing so, we will make as few (and as uncontroversial) linguistic assumptions as possible—in particular, we make no assumptions about the structure or semantics of Swiss German. We also present a few putative counterarguments and show that they are not seriously detrimental to our claim.

2. Some Swiss-German Data

Two facts about Swiss-German grammar are crucial to our argument. First, Swiss German uses case-marking (dative and accusative) on objects, just as standard German does; different verbs subcategorize for objects of different case. Second, Swiss German, like Dutch, allows cross-serial order for the structure of subordinate clauses.[4] Of critical importance is the fact that Swiss German requires appropriate case-marking to hold even within the cross-serial construction.

These linguistic claims are, however, stronger than the assump-

[3] Gazdar and Pullum [1982] provide a context-free grammar for the string set of Dutch, thus demonstrating its weak context-freeness, but they make no claim as to the linguistic motivation of the grammar.

[4] Though other orders are allowed as well, our argument is independent of such orders. See Section 4.2.

tions we need to show non-context-freeness. We will present some pertinent data below, later pinpointing exactly what claims we require for the proof. The sample subordinate clauses given here should be envisaged as preceded by the string "*Jan säit das*" ("Jan says that") or a similar precedent so as to form a complete sentence.

(1) *...mer em Hans es huus hälfed aastriiche*
...we Hans-DAT the house-ACC helped paint
...we helped Hans paint the house

Example (1) displays the cross-serial semantic dependencies found also in Dutch: *em Hans* is the object of *hälfed*, *es huus*, the object of *aastriiche*. Furthermore, correlated with this semantic dependency, there is a syntactic dependency between the pairs of constituents, namely, case-marking. The verb *hälfed* requires its *NP* object to be marked with dative case. A verb like *lönd*, which requires accusative case could appear in clauses like:

(2) *...mer de Hans es huus lönd aastriiche*
...we Hans-ACC the house-ACC let paint
...we let Hans paint the house

but not in

(3) *...* mer em Hans es huus lönd aastriiche*
...we Hans-DAT the house-ACC let paint
...we let Hans paint the house

Informants uniformly find this example ungrammatical and identify the case marking on *Hans* as the culprit. Similarly, since *aastriiche* requires an accusative object, the clause

(4) *...* mer de Hans em huus lönd aastriiche*
...we Hans-ACC the house-DAT let paint
...we let Hans paint the house

is also found to be ungrammatical.

This phenomenon of case marking across cross-serial verb constructions is quite robust, holding in quite complex clauses. For

example, the following triply embedded cross-serial clause is perceived as grammatical if and only if the case marking is correct.

(5) *...mer d'chind em Hans es huus lönd hälfe aastriiche*
...we the children-ACC Hans-DAT the house-ACC let help paint
...we let the children help Hans paint the house

(6) *...* mer d'chind de Hans es huus lönd hälfe aastriiche*
...we the children-ACC Hans-ACC the house-ACC let help paint
...we let the children help Hans paint the house

As further evidence of the robustness of the phenomenon, additional so-called raising verbs can occur between the string of *NP*s and the string of *V*s, e.g.:

(7) *...mer em Hans es huus haend wele hälfe aastriiche*
...we Hans-DAT the house-ACC have wanted help paint
...we have wanted to help Hans paint the house

(8) *...mer d'chind em Hans es huus haend wele laa hälfe aastriiche*
...we the children-ACC Hans-DAT the house-ACC have wanted let help paint
...we have wanted to let the children help Hans paint the house

3. A Non-Context-Freeness Argument

An argument for the weak non-context-freeness of Swiss German can be built from the foregoing data. On that basis we make the following minimal set of claims about the string set of Swiss German. Note that these claims are weaker than the analysis presented in the previous section.

Claim 1: Swiss-German subordinate clauses can have a structure in which all the *V*s follow all the *NP*s.

In particular, some sentences of the following schema are grammatical: *Jan säit das mer NP* es huus haend wele V* aastriiche* where the *NP*s are either *d'chind* or *em Hans* and the *V*s are either *laa* or *hälfe*. See sentences (7) and (8) for instances supporting this claim.

Claim 2: Among such sentences, those with all dative *NP*s preceding all accusative *NP*s, and all dative-subcategorizing *V*s preceding all accusative-subcategorizing *V*s are acceptable.

In particular, some sentences of the following schema are grammatical *Jan säit das mer (d'chind)* (em Hans)* es huus haend wele laa* hälfe* aastriiche.* Again, see sentences (7) and (8) for instances supporting this claim.

Claim 3: The number of *V*s requiring dative objects (e.g., *hälfe*) must equal the number of dative *NP*s (e.g., *em Hans*) and similarly for accusatives (*laa* and *d'chind*); note that this holds even if all the *V*s follow all the *NP*s.[5]

See sentences (6), and (12) through (22) for instances supporting this claim.

Claim 4: An arbitrary number of *V*s can occur in a subordinate clause of this type (subject, of course, to performance constraints).

Now, given any language L that satisfies these claims, we can take its image under the homomorphism f, where

[5]This claim holds, of course, only for those sentences in which the number of *NP*s equals the number of *V*s, as in all of the sample clauses presented here. Only sentences of this form are critical in the proof below, so that this weaker claim is still sufficient. Thus optionality of objects does not affect the proof and is not an issue here.

$$
\begin{aligned}
f(\text{“d'chind”}) &= a \\
f(\text{“em Hans”}) &= b \\
f(\text{“laa”}) &= c \\
f(\text{“hälfe”}) &= d \\
f(\text{“Jan säit das mer”}) &= w \\
f(\text{“es huus haend wele”}) &= x \\
f(\text{“aastriiche”}) &= y \\
f(s) &= z \text{ otherwise,}
\end{aligned}
$$

and then intersect the language $f(L)$ with the regular language $r = wa^*b^*xc^*d^*y$. According to the claims above, $f(L) \cap r = wa^mb^nxc^md^ny$, which is weakly non-context-free.[6] But since context-free languages are closed under homomorphisms and under intersection with regular languages [Hopcroft and Ullman, 1982, p. 130-135], the original language L, whatever it is, must also be weakly non-context-free. Now since our claims hold for Swiss German, the argument holds as well, and Swiss German is thus shown to be weakly non-context-free.[7]

As a trivial corollary, Swiss German is not strongly context-free either, regardless of one's view as to the appropriate structures for the language. Thus, we have an argument for the strong non-context-freeness of natural language that is not subject to the same frailty as the Dutch argument, i.e., its reliance on a linguistic motivation for its analysis of Dutch clause structure. Unlike the Dutch argument, ours does not mention, let alone hinge on, the constituent structure of the sentences in question or their semantics.

[6] This can be seen clearly by taking another image to remove the w, x and y, thereby yielding the standard example of a non-context-free language $a^mb^nc^md^n$ [Hopcroft and Ullman, 1979, p. 128].

[7] A similar argument showing the non-context-freeness of a fictitious language Dutch′ has been presented by Culy [1983].

4. Possible Counterarguments

The premises of the argument are quite explicit, namely the four claims presented above; counterarguments could be directed against any of them. We discuss several possibilities.

4.1. "The Data Are Wrong"

An argument can always be made that the grammaticality judgments expressed by our sample sentences are just wrong—that is, that the informants were mistaken about their own judgments or the transcriber simply misconstrued those judgments. This situation is, of course, hardly unique to this research, but pervades the linguistic method in general; it is especially problematic in the light of psychological research such as that of Rosenthal [1966]. It is the counterargument used against the "comparatives" argument [Gazdar and Pullum, 1982].

There being no adequate response to this objection, we will merely present details of our method in collecting the pertinent data and leave it to the reader to form an individual opinion. Four native Swiss-German speakers were interviewed separately, eliciting their grammaticality judgments on 62 Swiss-German clauses with varying word orders (disjoint, nested, cross-serial), depth of embedment, and lexical items. In an attempt to eliminate at least the most extreme of priming effects, the data were presented in a shuffled order. All four speakers were of the Zürich dialect of Swiss German, though one speaker claimed to have some Bernese traits in his dialect. (The Bernese dialect is freer than the Zürich in its constituent order.) The vast majority of examples (including all those presented in this paper except for (11)) showed unanimity of judgment among the speakers, and the phenomena came across as being surprisingly robust. It must be admitted, however, that the conclusions presented herein are not based on a controlled experiment. Such is usually and, for the most part, unavoidably the case in this area of linguistic research.

4.2. "Other Constituent Orders are Possible"

Claims 1 and 2 require that clauses allow a particular order in which all verbs follow all *NP*s and *NP*s and *V*s are "sorted" by case. Although we have noted that cross-serial orders may occur in Swiss-German subordinate clauses, other orders of constituents may also be permitted. Now, the mere fact that a certain subset of a language is non-context-free does not imply that the whole language is as well. This counterargument was effective against Postal's Mohawk argument, for instance, and the argument based on "respectively" constructions [Gazdar and Pullum, 1982].

Indeed, Swiss German does allow other constituent orders in relative clauses. For instance, the following examples are found to be grammatical:

(9) *...mer em Hans hälfed es huus aastriiche*
...we Hans-DAT helped the house-ACC paint
...we helped Hans paint the house

(10) *...mer em Hans es huus aastriiche hälfed*
...we Hans-DAT the house-ACC paint helped
...we helped Hans paint the house

and, depending on the particular dialect and context, even

(11) *...em Hans mer es huus hälfed aastriiche*
...Hans we the house helped paint
...we helped Hans paint the house

Similar examples can be found for the triply embedded examples.

However, the proof presented does not depend on the exclusion of orders other than the cross-serial. In fact, through intersection with the appropriate regular expression r, all sentences with other constituent orders or lexical items were removed from consideration. The proof is thus independent of the part of the language thereby abstracted. It is similarly immaterial whether or not the semantics of the construction is cross-serial, as the proof rests completely on the form of the sentences viewed as strings. (In fact, in Examples (9) through (11) above, the semantics are not strictly

cross-serial.) Finally, the argument does not hinge on any aspect of the *constituent structure* of the sentences whatsoever, since it is a purely formal stringset argument.

All that is critical is that no orders be allowed in which the case requirements of the verbs do not match the cases of the noun phrases (cf. Claim 3), but such clauses are found to be clearly ungrammatical whether cross-serial or not, e.g.,

(12) ...* *mer de Hans hälfed es huus aastriiche*
...we Hans-ACC helped the house-ACC paint
...we helped Hans paint the house

(13) ...* *mer em Hans hälfed em huus aastriiche*
...we Hans-DAT helped the house-DAT paint
...we helped Hans paint the house

(14) ...* *mer em Hans lönd es huus aastriiche*
...we Hans-DAT let the house-ACC paint
...we let Hans paint the house

(15) ...* *mer de Hans lönd em huus aastriiche*
...we Hans-ACC let the house-DAT paint
...we let Hans paint the house

(16) ...* *mer de Hans es huus aastriiche hälfed*
...we Hans-ACC the house-ACC paint helped
...we helped Hans paint the house

(17) ...* *mer em Hans em huus aastriiche hälfed*
...we Hans-DAT the house-DAT paint helped
...we helped Hans paint the house

(18) ...* *mer em Hans es huus aastriiche lönd*
...we Hans-DAT the house-ACC paint let
...we let Hans paint the house

(19) ...* *mer de Hans em huus aastriiche lönd*
...we Hans-ACC the house-DAT paint let
...we let Hans paint the house

(20) ...* *mer de Hans haend wele hälfe es huus aastriiche*
...we Hans-ACC have wanted help the house-ACC paint
...we have wanted to help Hans paint the house

(21) ...* *mer d'chind lönd de Hans hälfe es huus aastriiche*
...we the children-ACC let Hans-ACC help the house-ACC paint
...we let the children help Hans paint the house

(22) ...* *mer d'chind de Hans es huus lönd hälfe aastriiche*
...we the children-ACC Hans-ACC the house-ACC let help paint
...we let the children help Hans paint the house

Thus, additional permitted orders of constituents do not provide a counterargument to our first two claims, or our conclusion.

4.3. "Case Is Not Syntactic"

An argument could be put forth that Claim 3 is in error. Case agreement, one might argue, need *not* hold for these sentences to be *syntactically* correct; case agreement, one would then hold, is actually extrasyntactic, perhaps even semantic. This type of argument was used against both the "respectively" non-context-freeness argument and the argument based on the digits of π [Gazdar and Pullum, 1982].

Clearly, the burden of proof is on the proponent of this straw man to furnish some evidence for the radical claim that case marking in Swiss German is a purely extrasyntactic or semantic notion. It would need to be demonstrated that the case requirements of verbs are completely predictable from their meanings. In particular, it is not sufficient to note that the case marking on *NP*s provides information as to the semantic role played by the *NP* in a clause.

Certainly, the native informants did not find the starred clauses above semantically anomalous, but ungrammatical. No consistent semantic distinction between raising verbs requiring a dative ob-

ject and those taking an object in the accusative case seem forthcoming, nor do clear distinctions between the meanings of dative versus accusative *NP*s independent of context. Finally, in related languages, e.g., German and Dutch, case is widely considered a purely syntactic phenomenon.

4.4. "Clauses are Bounded in Size"

Finally, Claim 4 could be rejected. Much beyond triple embedding of clauses, judgments get weaker (though it should be noted that the judgments on Clause (5) and the even more deeply embedded Clause (8) did not seem to be on the margin of performance bounds). One could argue that the phenomenon of cross-serial clause structure is bounded by, say, five embeddings or, to be more generous, one hundred. In either case, the language with bounded cross-seriality would be context-free, regardless of case-marking properties.

Down this path lies tyranny. Acceptance of this argument opens the way to proofs of natural languages as regular, nay, finite. The linguist proposing this counterargument to salvage the context-freeness of natural language may have won the battle, but has certainly lost the war.

5. Conclusion

Using a particular construction of Swiss German, the cross-serial subordinate clause, we have presented an argument providing evidence that natural languages can indeed cross the context-free barrier. The linguistic assumptions on which our proof rests are small in number and quite weak; most of the proof is purely formal. In fact, the argument would still hold even if Swiss German were significantly different than it actually is, allowing many more constituent orders, cases and constructions, and even if the meanings of the sentences were completely different.

What has *not* been shown by this argument is equally impor-

tant to keep in mind. By proving the non-context-freeness of the language of the Swiss-German competence grammar, we have still not demonstrated that natural languages are impossible, or even difficult, to parse. Both the Dutch and Swiss-German constructions are linear-parsable, and, were they not so in theory, performance constraints might well make them so. We have not demonstrated that powerful grammar formalisms with context-sensitive or even the weaker indexed power are essential for describing natural language. Indeed, the difficulty of finding evidence for the non-context-freeness of natural language remains a challenge and mystery.

In a more speculative vein, we believe that, though the search for tight formal constraints on grammars and restrictive mathematical properties of natural languages (in the spirit of the context-free hypothesis) is a worthy goal, the present research may be a clue leading in a slightly different methodological direction. It raises the possibility that the most revealing account of a natural language may be one in which the formalism describing the competence grammar is powerful, well beyond context-free power, but where the learning, parsing, and/or generation mechanisms provide the constraints that mutually allow learnability, parsability, and generability. The search for formalism restrictions should therefore be accompanied by research on precise models of language mechanisms, which may one day lead to a resolution of the Swiss-German paradox and challenge—to find theories that are powerful enough to yield revealing accounts of complex data, yet restrictive enough to be explanatory in form.

References

[Bresnan, *et al.*] Bresnan, J., R.M. Kaplan, S. Peters, and A. Zaenen, 1982: "Cross-Serial Dependencies in Dutch."

[Chomsky, 1963] Chomsky, N., 1963: "Formal Properties of Grammars," in *Handbook of Mathematical Psychology,* volume II, R.D.

Luce, R.R. Bush, and E. Galanter, eds. (New York: John Wiley and Sons), pp. 323-418.

[Culy, 1983] Culy, C.D., 1983: "An Extension of Phrase Structure Rules and its Application to Natural Language," Master's thesis, Stanford University, Stanford, California (May).

[Gazdar, 1982] Gazdar, G., 1982: "Phrase Structure Grammar," in P. Jacobson and G.K. Pullum, eds., *The Nature of Syntactic Representation*, (Dordrecht, Holland: D. Reidel Publixhing Company).

[Gazdar and Pullum, 1982] Gazdar, G. J. M. and G. K. Pullum, 1982: "Natural Languages and Context-Free Languages," *Linguistics and Philosophy*, Volume 4, Number 4, pp. 469-470.

[Higginbotham, 1984] Higginbotham, J., 1984: "English is not a Context-Free Language," *Linguistic Inquiry*, Vol. 15, pp. 119-126.

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

[Joshi, 1983] Joshi, A.K., 1983: "How Much Context-Sensitivity is Required to Provide Reasonable Structural Descriptions: Tree Adjoining Grammars," to appear in D. Dowty, L. Karttunen, and A. Zwicky, eds., *Natural Language Processing: Psycholinguistic, Computational, and Theoretical Perspectives*, (Cambridge, England: Cambridge University Press).

[Rosenthal, 1966] Rosenthal, R., 1966: *Experimenter Effects in Behavioral Research*, (New York, New York: Appleton-Century-Crofts).

[Thompson, 1983] Thompson, H., 1983: "Crossed Serial Dependencies: A Low-Power Parseable Extension to GPSG," *Proceedings of the 21st Annual Meeting of the Association for Computational Linguistics*, Massachusetts Institute of Technology, Cambridge Massachusetts (15-17 June).