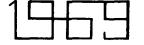On the Preservation of Context-Free Languages in
A Level-Based System

- Jacob Mey -

INTERNATIONAL CONFERENCE

ON

C*O*MPUTATIONAL L*I*N*G*UISTICS

⇓

C O L I N G

1969

ON THE PRESERVATION OF CONTEXT-FREE LANGUAGES IN

A LEVEL-BASED SYSTEM

Jacob Mey

The University of Texas at Austin
Austin, Texas 78712
U.S.A.

Abstract *

In this paper, a recently proposed level-oriented
model for machine analysis and synthesis of natural
languages is investigated. Claims concerning the
preservation of context-free (CF) languages in such
a system are examined and shown to be unjustified.
Furthermore, it is shown that even a revised version
of the model (incorporating some recent discoveries)
will not be CF-preserving. Finally, some theoretical
implications of these findings are explored: in par-
ticular, claims of greater naturalness and the ques-
tion of recursivity.

Résumé *

Dans ce travail, on envisage un modèle récent de
synthèse et d'analyse automatiques de langues naturelles,
orienté vers la notion de "niveau" linguistique. On
examine un postulat selon lequel les langages context-
free seraient stables dans un tel système; cette notion
s'avère incorrecte. De plus, on montre comment même une
version modifiée du modèle, incorporant certaines dé-
couvertes récentes, ne conserve pas le caractère con-
text-free de ces langages. Enfin, on explore l'impor-
tance théorique de ces résultats; en particulier, on
examine l'avantage supposé d'un modèle dit plus "natu-
rel", et la question de la récursivité des langues na-
turelles.

---

1. Level-Oriented Systems in Computational Linguistics

A new model for computational linguistic perfor-
mance has recently been proposed by the Czechoslovak
group of workers at Charles University, Prague, under
the direction of P. Sgall (1, 2). This model has sig-
nificant theoretical implications, since it offers
an alternative to transformationally based solutions
of the problems encountered in automatic syntactic
analysis , and consequently, to the transformational
model itself. In particular, the new model claims to
compete favorably with the transformational one in
generative power and structural characterization of
sentences.

Central to this model is the notion of a "multi-level"
or "stratified" grammar. The generation of a sentence
at the highest ("deepest") level of the grammar pro-
ceeds by a set of context-free (CF) rules; the output
of these rules is then transduced to lower levels by
a series of pushdown store automata. The output of the
final transduction is some "surface" representation
of the sentence to be generated.

2. CF-Preservation under Transduction

The whole system described in the preceding section is
said to be "weakly equivalent to a CF phrase structure
grammar" (1:148, 221). This assertion is claimed to
derive from theorems formulated by N. Chomsky and R.
J. Evey (3; 4), which maintain that the output of a
pushdown store transducer (pdt) is equivalent to the
set of CF languages (1:109; 2:2.2.5).
However, the original theorems about this equivalence
concerned pushdown store acceptors (pda), not trans-
ducers (pdt as defined by Ginsburg (5:102)); for pdt
(as used by Sgall and his group), CF-preservation is
known not to obtain in general (5:104).
The condition under which pdt will preserve CF langua-
ges is stated by Ginsburg in Th. 3.5.1. (5:104):
given a pda M, a pdt S, L = T(M),
S(L) is CF iff M and S are associated, i.e. S is ob-
tained by adding outputs to M, the pda that accepts L.
For the case of the Czechoslovak "battery" of trans-
ducers, this condition could actually be fulfilled,
although the authors never say so explicitly. I am

referring to their extra condition on the system,
as formulated in "the existence of inverse automata
for the single levels of the pushdown part [of the
grammar, JM] " (1:146). Hence the impreciseness of
the Czechoslovak proposal may amount to no more than
a matter of incomplete formulation.

In a recent article, however, Ginsburg and Rose (6)
have shown that the earlier theorem on which they
based CF-preservation conditions for pdt is false.
(And, by the same token, so are the original theorem
of Evey's (4:2.6.6) and an earlier theorem by Ginsburg
and Rose (7:3.2)). According to the 1968 revised ver-
sion of the latter theorem by Ginsburg and Rose (6:3.
2a), CF-preservation is made dependent upon an addi-
tional condition on the pdt, namely, that of restrict-
ing its output to those strings that are produced by
the device when it ends up in an accepting (or final)
state. In all other cases, the language generated by
the pdt will not be equivalent to the set of CF lan-
guages, but simply constitute a recursively enumer-
able set.

3. Practical Consequences for the Prague System

For the case under consideration, the new insight re-
ferred to in the preceding section has two signifi-
cant consequences:

first, the worries of the Czechoslovak group to ensure
CF preservation may well have been in vain, unless
the new condition can be incorporated into their sys-
tem. Otherwise, a device that is practically equiva-
lent to a Turing machine is not very exciting to work
with in computational linguistic theory or its imple-
mentation.

Second, one of the advantages inherent in the use of
CF-preserving transducers is the guaranteed existence
of a whole bevy of working recognition routines (e.g.,
the Cocke-Robinson algorithm, the parser developed
by Kay, or the predictive analyzer by Kuno, etc.)
This advantage becomes illusory if the pdt battery
produces a recursively enumerable language, i.e. one
that cannot be guaranteed to be recognized by a CF-
recognition routine.

If we think of the Czechoslovak system as part of a

machine translation proposal, where the route from
source to "interlingua" consists of essentially the
same flow (but in the opposite direction) as that
from "interlingua" to target language, it appears
that the consequences of an incorporation of the
Ginsburg-Rose adjustment are far-reaching. Such an
incorporation could take place in two ways: either
one could check the output of a particular (i-th)
device, transducing from a higher (i-th) to a lower
(i - 1th) level, or from a lower (i-th) to a higher
(i + 1th) level, to see whether or not this output
corresponds to an accepting state (where "higher"
and "lower" are understood to refer to deeper and
more superficial structures respectively); or, al-
ternatively, a built-in checking device could prevent
output from being generated unless the transducer
reached an accepting state after reading the input
string. So far, Sgall and his group have not sugges-
ted ways to handle this problem.

Quite another matter is that the Prague group's pro-
posal to let the output of the i-th device be a pro-
per subset of the input of the i + 1th, i - 1th

transducer respectively, does not seem to be fruitful,
or even feasible. Naturally, the first question that
arises is: what about the remaining input, where does
it all come from? It is certainly true, as the authors
remark (2:2.2.5), that "the output language of the
whole description is not necessarily context-free".
As shown above, it simply never is under the given
conditions. Hence the reason given in the rest of the
quote is trivial: "since ... it is only a proper sub-
set of the context-free languages of the last pushdown
transducer" (ibid.). While it is always possible to
take a proper subset of a CF language and obtain a
language that is not CF, or maybe not even regular,
that clearly is not the point here. The authors intend
their output language to be CF, for reasons like the
ones mentioned above. As long as it can be shown (as
I have done here) that the system in no case is (even
weakly) equivalent to a CF grammar, the question of
the restrictions on the input to the subsequent trans-
ducers is, of course, irrelevant to the CF character
of the system as a whole.

4. Some theoretical implications

The proponents of the Prague model have repeatedly
asserted that their system has certain advantages
over other models of linguistic performance and, in
particular, that their grammar is superior (or at
least equivalent) to transformational grammar.
The claim that the level-based model is superior to
others because of easy CF-recognition has been dispro-
ved in the preceding sections. Another claim, that of
greater naturalness inherent in the level-oriented
model, is also often made (sometimes implicitly by
reference to the model's stance in time-honored lin-
guistic tradition). It should be observed, however,
that such a claim does not concern the formal charac-
ter of any system. As Chomsky has pointed out (in his
discussion of Fillmore's case theory (8:14-16)), it
is vacuous to discuss different formal systems in
terms of which is the more "direct" representation
of natural language; unless a formal system is inter-
preted, it simply cannot be compared to another one
for "directness" of expression.
But how about transformational grammar itself with

respect to sentence recognition? It has been known
for a long time that there is no way of establishing
a universal automatic recognition procedure for a
transformational grammar that is not in some ways
restricted. This was precisely the Prague group's
motivation for proposing their system as a (superior)
alternative to TG. Now that their claims have been
de-substantiated on theoretical grounds, it may seem
like a meagre consolation to the Prague group that TG
itself is in the same boat, theoretically speaking.
In an important recent study, Peters and Ritchie have
demonstrated that a context-sensitive (CS) based
transformational grammar, unless restricted in some
respects, generates a recursively enumerable language,
and conversely, for any recursively enumerable language
there is a CS based TG that will generate it (9:4.1).
As the authors point out (ibid.:33), it is imperative
to set out and find conditions under which a TG will
generate only recursive languages. One such condition
would be to restrict the base of a TG to be CF; however,
this will still not guarantee recursivity (9:4.2).
This is precisely the problem that the Czechoslovak
workers will have to solve in order to make their system
viable and to validate their claims.

References

(1) P. Sgall, <u>Generativní Popis Jazyka a Česká Deklinace</u>, Prague 1967.

(2) P. Sgall & al., <u>A Functional Approach to Syntax</u>, New York 1969 (in press).

(3) N. Chomsky, "Formal Properties of Grammars", in: R. Luce, R. Bush & E. Galanter (edd.), <u>Handbook of Mathematical Psychology</u>, II, New York 1963, pp. 323-418.

(4) R. Evey, <u>The Theory and Application of Pushdown Store Machines</u>, Cambridge, Mass., 1963. (doct. diss.)

(5) S. Ginsburg, <u>The Mathematical Theory of Context-Free Grammars</u>, New York 1966.

(6) S. Ginsburg & G. Rose, "A Note on Preservation of Languages by Transducers", <u>Inf. Contr.</u> 12(1968):549-552. (See also (10), below)

(7) S. Ginsburg & G. Rose, "Preservation of Languages by Transducers", <u>Inf. Contr.</u> 9(1966): 153-176. (See also (10), below)

(8) N. Chomsky, <u>Deep</u> <u>Structure</u>, <u>Surface</u> <u>Structure</u>, <u>and</u> <u>Semantic</u> <u>Interpretation</u> (mimeographed), M.I.T. 1968

(9) S. Peters & R. Ritchie, <u>On</u> <u>the</u> <u>Generative</u> <u>Power</u> <u>of</u> <u>Transformational</u> <u>Grammars</u> (mimeographed), Tech. Report CSci. 69-2-3-, The University of Washington, Seattle 1969.

(10) J. Mey, Review of Ginsburg & Rose (6), <u>Comp</u>. <u>Rev</u>. 10(May 1969), in press. (This review contains a succinct statement of the contents of (6) and (7), as well as a historical footnote on the authorship of the main theorem of Section 2 of (7) and its revised form as presented in (6)).