[Mechanical Translation and Computational Linguistics, vol.9, no.1, March 1966]

A Procedure for Morphological Encoding

by P. H. Matthews, Department of Linguistic Science, University of Reading, England

A *finite-state machine is described which will control the derivation of Italian verb forms, including proper stress placement, given an appropriate dictionary and set of grammatical rules.*

I. Introduction

In many languages a word may be identified, on the syntactic level, by a single vocabulary element or *lexeme* and a single term from each of a set of closed grammatical categories.[1] For example, the Italian verb form *canterá* (possible translation: "he will sing") may be identified, on the one hand, by a vocabulary element which we symbolize in the form CANTARE and, on the other, by the terms "Future" (Fu) and "non-Past" (non-Pa) from the categories TENSE^{a} and TENSE^{b}, the term "Indicative" (Ind) from the category MOOD, and the terms "third Person" (3) and "singular" (sg) from the categories PERSON and NUMBER. (The categories TENSE^{a} [Future and non-Future] and TENSE^{b} [Past and non-Past] are postulated on morphological grounds: this proposal is tentative but may well have syntactic and semantic justification. The various forms discussed in this paper are customarily displayed in paradigms; for example, see Reynolds [1962] for the paradigms of MANDARE, a verb of the same class as CANTARE, and STARE [see below]. A less "traditional" account of Italian morphology, though inevitably dated, can be found in Hall [1949].) Future, Indicative, etc., are interpreted here as properties (we will call them *morphosyntactic properties)* of the word concerned. Thus *canterá,* we will say, is that form of the vocabulary element CANTARE which has all and only the morphosyntactic properties non-Past, Future, Indicative, third Person, and singular. For such a syntactic representation we will employ the notation

$$\text{CANTARE}_{\text{Fu, non-Pa, Ind, 3, sg}}$$

(following the traditional verbalization "the third singular Future non-Past Indicative of CANTARE").

For the same languages, the realization of a word (expressed as a string of letters, a string of morphophonemes, and so on) may be derived from the root of the relevant vocabulary element by a finite sequence of morphological operations. Thus the form *canterá,* given that the root of CANTARE has the form *cánt,* might be derived by the suffixation of *er (cánt → cánter),* the suffixation of *a (cánter→ cántera),* and the shifting of the stress (symbolized by the acute accent) from the first vowel to the third. Each choice of operation may be determined by either or both of the following factors: first, by some particular subset of the relevant morphosyntactic properties and, second, by the morphological class to which the vocabulary element involved must be assigned. Thus the *a*-suffix in *canterá* is selected for all words with the properties Future, non-Past, third Person, and singular; contrast *canteró* (CANTARE$_{\text{FU, non-Pa, Ind, 1[st Person], sg}}$), *canto* (CANTARE$_{\text{non-Fu, non-Pa, Ind, 3, sg}}$), etc. The *er*-suffix, on the other hand, is not only restricted to words with the property Future but is further restricted to a class of vocabulary elements that has CANTARE, but not VEDERE, PARTIRE, etc., among its members. Contrast *vedrá* (VEDERE$_{\text{Fu, non-Pa, Ind, 3, sg}}$), *partiró*(PARTIRE$_{\text{Fu, non-Pa, Ind, 1, sg}}$), and so forth. The purpose of this paper is to describe a procedure which, given the syntactic representation of some particular word, will determine (from an appropriate dictionary and set of grammatical rules) that precise sequence of operations by which its realization is derived. The form of rule required will be introduced in Section II. The procedure itself will be presented in Section III.

II. Inflectional Rules

Let us begin by considering the problem from a slightly different angle. It is clearly possible to devise a finite-state machine that will generate all and only those sequences of operations that are required for the word forms of a given language. A part of such a machine is shown in Figure 1. The sequences which this will generate are those required for the Future forms both of CANTARE and of the partly irregular verb STARE, in Italian. In Figure 1 we take account of all the stresses, not merely of those that happen to be indicated by the orthography. For example, the sequence of operations

[Suffix] *er,* SFV [Stress Following Vowel], [Suffix] *e,* [Suffix] *bbe*

(the machine terminates in s_4 after passing through s_1 and s_2) is intended to yield the form *canterébbe;* by the first operation *cánt → cánter,* by the third and second *cánter → canteré,* and by the fourth *canteré → canterébbe.* Likewise, the sequence

ar, SFV, *e,* SPV [Stress Preceding Vowel], *mo*

[1] Preliminary versions of this paper were presented to a conference at the RAND Corporation in August, 1963, and to the Mechanolinguistics Colloquium at Berkeley in May, 1964; I am grateful for comments and assistance received on both occasions. The model involved has since been discussed in greater detail by Matthews (1965). The illustrations in this paper are intended for illustration only; they should not be taken as a serious contribution to the description of Italian.

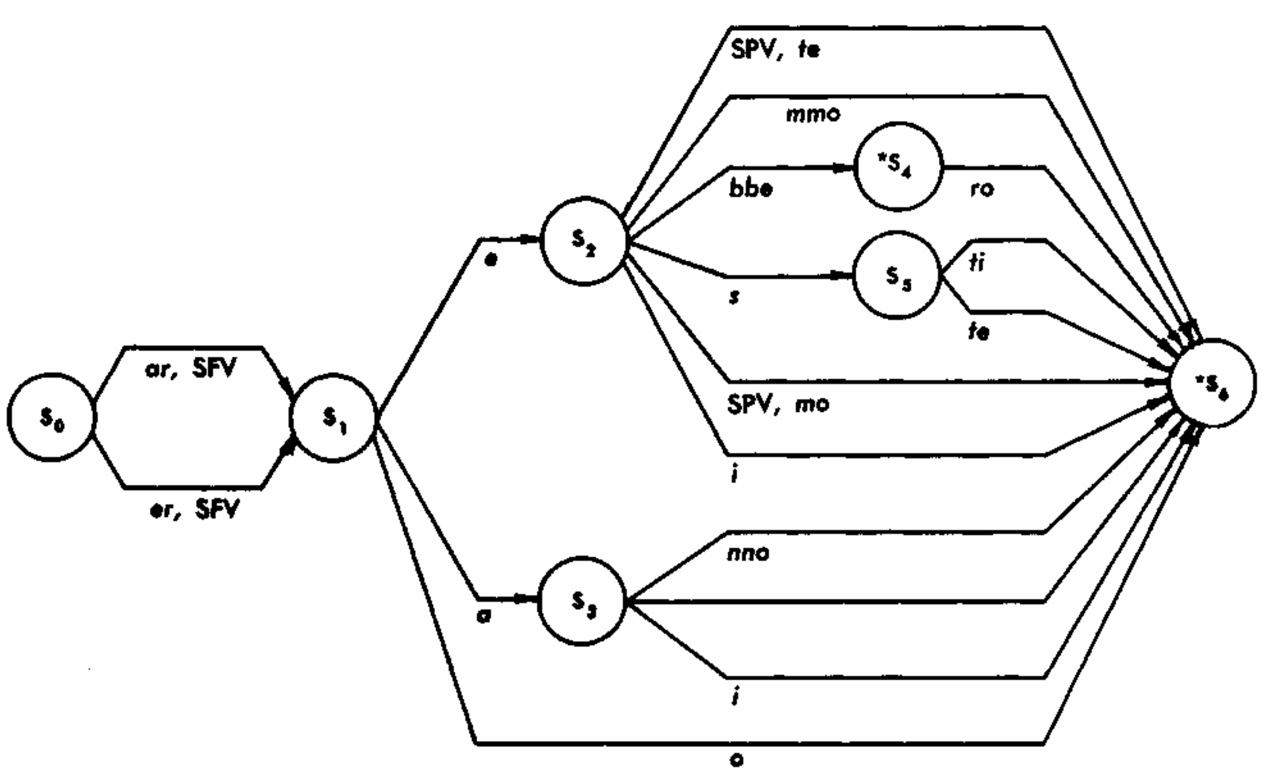


FIG. 1.—The finite-state machine. SFV, stress following vowel; SPV, stress preceding vowel. States are numbered s_1, s_2, etc., and final states are marked with an asterisk.

(the machine terminates in s_6 after passing through s_1 and s_2) is intended to yield the form *starémo;* by the first operation a form *star* is derived from a root *st,* by the third and second *star* → *staré,* by the fourth *staré* → *staré,* and by the fifth *staré* → *starémo.* (SPV and SFV are understood to move the stress, if necessary, to the vowel indicated. In the case of SPV, it is moved to the last vowel in the current operand; given *canteré* as the operand [which would result from the application of *er,* SFV, and *e],* SPV would apply vacuously to yield *canteré.* In the case of SFV, on the other hand, the application of a similar operation is held over until subsequent suffixation has added a further vowel to the operand. Thus, given the root *cánt* as the initial operand, the sequence *er,* SFV, *a* will apply as follows: first by *er, cánt* → *cánter;* second, *cánter* → *cántera* by *a,* SFV being held over; third, SFV applies to yield *canterá.* In this restricted illustration SPV always applies vacuously; however, this represents an extension, to the Future forms, of rules that apply non-vacuously to handle *cantiámo, cantaváte,* etc.; see rules 13 and 15 in the sample below.)

Such a machine may well be adequate for some purposes; its disadvantage, however, is that it fails to indicate which particular sequence of operations is appropriate to which particular word. Figure 1 may generate the sequences required for *canterébbe, starémo,* etc., but it does not indicate that *canterébbe* is the realization of $\text{CANTARE}_{\text{Fu, Pa, Ind, 3, } sg}$ or that *starémo* is the realization of $\text{STARE}_{\text{Fu, non-Pa, Ind, 1, pl[ural]}}$. Our problem may accordingly be represented as follows. How should we specify, for a machine of this kind, the set of words for which each transition must be selected? How do we indicate, for example, that of the transitions from s_0 to s_1 one is appropriate to STARE and the other to CANTARE?

Our solution requires, in the first place, that each state should be labeled with an *index symbol.* For the single initial state (s_0 in Fig. 1) we will employ the index symbol *R*; *R* may be interpreted, in linguistic terms, as the set of all roots in the language. For each final state (s_4 and s_6) the label will be one of a set of *form-class* symbols, in this case a symbol *V* which may be interpreted, in linguistic terms, as the set of all verb forms. Of the remaining states in Figure 1, s_1 will be labeled with the symbol *C,* s_2 and s_3 with the symbol *S*, and s_5 with the symbol *M*; it may help to interpret these as classes of stems, for example, the stem *canteré* in *canterébbe,* etc., or the stem *starés* in *starésti* and *staréste.* Given such index symbols, each transition may be represented by a rule with one optional and two obligatory components. The first component, which we will call the *reference* component, is obligatory; its form is as follows:

$$[I_{q1,\, q2,\, \ldots\ldots\, qn}],$$

where *I* is the label of the state resulting from the transition and *{*q_1*,* q_2*,. . ., q*$_n$*}* is a set of zero or more morphosyntactic properties. The second component, which we will refer to as the *limitation,* is optional; where a rule has such a component it will be of the form *A*, where *A* is a class of vocabulary elements. Finally, the third component, which we will refer to as the *formation* component (in preference to "representation" or "representation component" in Matthews [1965]), is of the form

$$o_1, o_2 \ldots, o_n, B,$$

where o_1, o_2*, . . . , o*$_n$ is a sequence of zero or more morphological operations and where *B* (which we will refer to as the base component) is a further expression of the form

$$[I_{q1, q2, \cdots\cdot qn}],$$

I being, in this case, the label of the state preceding the transition and {q_1, q_2,. . . , q_n*}* being a further set of zero or more morphosyntactic properties. An example would be the rule

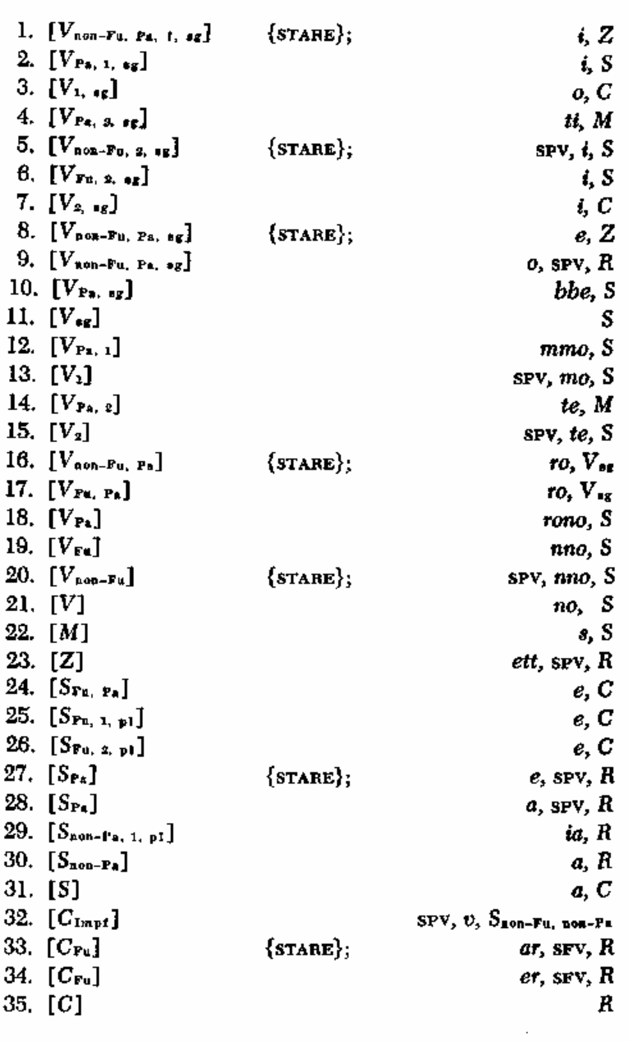
[C_{Fu}] {STARE}; *ar, SFV, R,*

which corresponds, in the set of rules presented below, to the transition between s_0 and S1 which is uppermost in Figure 1. Another would be a rule

[$V_{Fu, \text{non-Pa}, 3, pl}$] *ro,* V_{sg},

(compare rule 17 below) which might correspond to the transition between s_4 and s_6. The first of these examples has a limitation (see above) which indicates that it is valid only for members of the set {STARE}. The second has no such limitation and might be verbalized as follows: for all verbs, the Future, non-Past, third Person plural is derived from the corresponding singular form by the suffixation of *ro*.

Let us now introduce a more extended illustration. The rules below will handle all the Indicative forms of STARE and CANTARE, including those generated in Figure 1. Of the transitions in Figure 1 those from s_0 to s_1 correspond to rules 33 and 34; those from s_1 to s_2 and s_3 to rules 24-26 and 31; that from s_1 to s_6 to 3; that from s_2 to s_4 to 10; that from s_2 to s_5 to 22; those from s_2 to s_6 to 15, 12, 13, and 6; those from s_3 to s_6 to 19, 11, and again 6; that from s_4 to s_6 to 17; and those from s_5 to s_6 to 4 and 14. (However, most of these rules are generalized to cover additional cases.) Note that the procedure in Section III will interpret these rules as ordered; for example, rule 2 will apply only in those cases not covered by rule 1, and rule 3 only in those cases not covered by 1 and 2. Where the derivations differ from one verb to the other (e.g., in the cases handled by 8 and 9), the rule for STARE is written first and the rule for CANTARE (to be precise, for all relevant verbs except STARE) later. Note also, in rule 32, that we have retained the traditional term "Imperfect" (Impf); for example, *cantáva* is the realization of CANTARE$_{\text{Impf, Ind, 3, sg}}$. This may be thought of as a third member of the category TENSE^b; unlike Past and non-Past, it entails a "neutralization" of the distinction within TENSE^b.



1. [$V_{\text{non-Fu, Pa, 1, sg}}$]	{STARE};	*i, Z*
2. [$V_{\text{Pa, 1, sg}}$]		*i, S*
3. [$V_{\text{1, sg}}$]		*o, C*
4. [$V_{\text{Pa, 2, sg}}$]		*ti, M*
5. [$V_{\text{non-Fu, 2, sg}}$]	{STARE};	SPV, *i, S*
6. [$V_{\text{Fu, 2, sg}}$]		*i, S*
7. [$V_{\text{2, sg}}$]		*i, C*
8. [$V_{\text{non-Fu, Pa, sg}}$]	{STARE};	*e, Z*
9. [$V_{\text{non-Fu, Pa, sg}}$]		*o,* SPV, *R*
10. [$V_{\text{Pa, sg}}$]		*bbe, S*
11. [V_{sg}]		*S*
12. [$V_{\text{Pa, 1}}$]		*mmo, S*
13. [V_{1}]		SPV, *mo, S*
14. [$V_{\text{Pa, 2}}$]		*te, M*
15. [V_{2}]		SPV, *te, S*
16. [$V_{\text{non-Fu, Pa}}$]	{STARE};	*ro,* V_{sg}
17. [$V_{\text{Fu, Pa}}$]		*ro,* V_{sg}
18. [V_{Pa}]		*rono, S*
19. [V_{Fu}]		*nno, S*
20. [$V_{\text{non-Fu}}$]	{STARE};	SPV, *nno, S*
21. [*V*]		*no, S*
22. [*M*]		*s, S*
23. [*Z*]		*ett,* SPV, *R*
24. [$S_{\text{Fu, Pa}}$]		*e, C*
25. [$S_{\text{Fu, 1, pl}}$]		*e, C*
26. [$S_{\text{Fu, 2, pl}}$]		*e, C*
27. [S_{Pa}]	{STARE};	*e,* SPV, *R*
28. [S_{Pa}]		*a,* SPV, *R*
29. [$S_{\text{non-Pa, 1, pl}}$]		*ia, R*
30. [$S_{\text{non-Pa}}$]		*a, R*
31. [*S*]		*a, C*
32. [C_{Impf}]		SPV, *v,* $S_{\text{non-Fu, non-Pa}}$
33. [C_{Fu}]	{STARE};	*ar,* SFV, *R*
34. [C_{Fu}]		*er,* SFV, *R*
35. [*C*]		*R*

III. Description of the Procedure

A suitable encoding procedure may be summarized by the flow chart in Figure 2. It falls into four sections (Boxes A1-A2, B1-B6, C1-C2, and D1-D8), which may be described as follows.

SECTION A

The procedure encodes one word at a time. As a first step, the relevant lexeme symbol is entered in a location LEXEME, and the accompanying morphosyntactic properties form the first entries in a block SUBSCRIPT (Box Al). Thus, for the word realized by *canterébbero,* LEXEME and SUBSCRIPT will read:

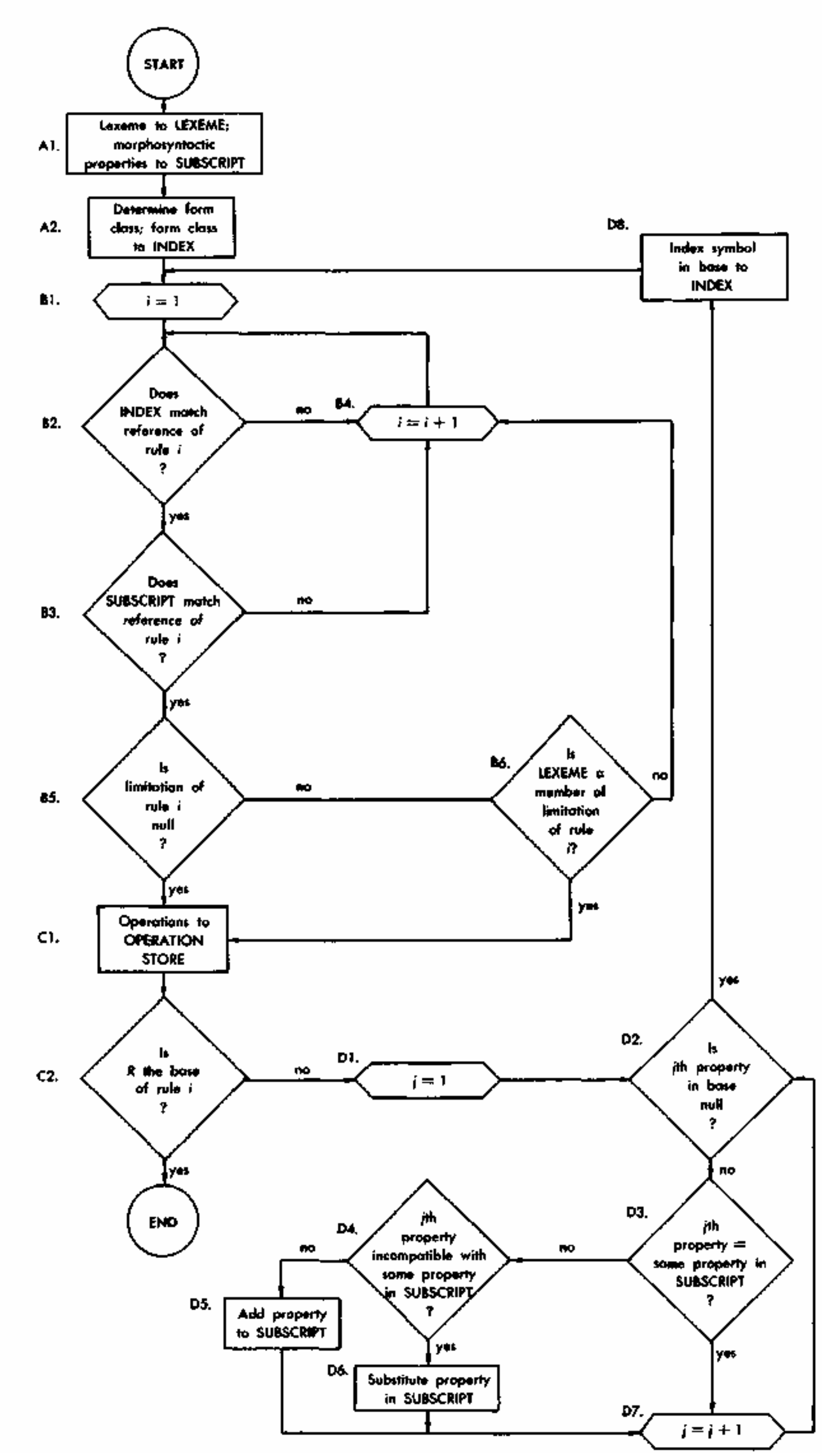

START
A1.
Lexeme to LEXEME; morphosyntactic properties to SUBSCRIPT
A2.
Determine form class; form class to INDEX
D8.
Index symbol in base to INDEX
B1.
i = 1
B2.
Does INDEX match reference of rule i ?
no
B4.
i = i + 1
yes
B3.
Does SUBSCRIPT match reference of rule i ?
no
yes
B5.
Is limitation of rule i null ?
no
B6.
Is LEXEME a member of limitation of rule i?
no
yes
yes
C1.
Operations to OPERATION STORE
yes
C2.
Is R the base of rule i ?
no
D1.
j = 1
D2.
Is jth property in base null ?
no
yes
END
D4.
jth property incompatible with some property in SUBSCRIPT ?
no
no
D3.
jth property = some property in SUBSCRIPT ?
D5.
Add property to SUBSCRIPT
yes
D6.
Substitute property in SUBSCRIPT
yes
D7.
j = j + 1

FIG. 2.—Encoding procedure. Procedure represented by flow chart assumes that search cannot fail—which, in the case of an adequate set of rules and an acceptable input, I suppose to be true.

LEXEME CANTARE

SUBSCRIPT Pa
Fu
Ind
3
pl

The procedure then determines the appropriate form class (e.g., as part of a dictionary lookup for the lexeme CANTARE) and enters this in a location INDEX (A2). Continuing with the same example, INDEX will then read:

INDEX V

SECTION B

The next routine refers to these entries to identify a particular inflectional rule; this will correspond to one of the final transitions (e.g., the transition from s_4 to s_6) in a machine of the type shown in Figure 1. The rule concerned must meet three conditions. First, the current entry in INDEX must match the index symbol which forms part of its reference component (B2); thus if V is entered in INDEX, all of rules 22-35 are excluded. Second, the morphosyntactic properties referred to by its reference component must form a subset of the current entries in SUBSCRIPT (B3); if SUBSCRIPT reads as above, this excludes all of rules 1-11 *(inter alia* because singular is not one of the entries), 12 and 13, etc., but does not exclude 17-19. Third, the rule either must have no limitation (B5), or, if it has a limitation, then the morphological class referred to must have the lexeme entered in LEXEME as a member (B6); normally, this would presuppose a dictionary lookup for the lexeme concerned. Since inflectional rules are ordered (see Sec. II, above), the procedure makes a continuous pass (Bl and B4) until a rule that meets all three conditions has been located. With the above entries in LEXEME, INDEX and SUBSCRIPT, the first to do so will be rule 17.

SECTION C

The third routine examines the formation component of the rule identified in Section B.

1. First, the operations listed (if any) are added to the existing entries (if any) in a block OPERATION STORE (C1): thus if rule 17 was the first rule in question, the first entry in OPERATION STORE would read:

OPERATION STORE *ro*

This block will be treated as a pushdown. New entries will be made above existing entries; furthermore, the operations listed in any one formation component will be entered in reverse order. Let us suppose, for instance, that the rules identified in subsequent cycles are rules 10, 24, and 34. Of these, 10 and 24 list one operation each; the operations concerned will therefore be entered in OPERATION STORE as follows:

OPERATION STORE *e*
bbe
ro

Rule 34, on the other hand, mentions two: successively *er* and SFV. Entering the second of these first, OPERATION STORE will accordingly be extended to read:

OPERATION STORE *er*
SFV
e
bbe
ro

It will be seen that the contents of this block, reading from top to bottom, would then consist of the sequence of operations required (see Fig. 1) for the derivation of *canterébbero.*

2. At this point, the procedure will either terminate or it will pass to another cycle. If the base component consists of the single symbol *R,* it terminates (C2); the rule concerned would correspond to one of the initial transitions (e.g., to one of the transitions from s_0 to s_1) in a diagram such as Figure 1. If not, it proceeds to Section D.

SECTION D

The fourth section revises the entries in INDEX and SUBSCRIPT in preparation for the next pass through the grammar. For this purpose, it too refers to the base component of the rule found in Section B.

1. The entries in SUBSCRIPT are considered first. If no morphosyntactic properties are mentioned in the base component (D2), SUBSCRIPT is unchanged. Otherwise the procedure takes each property in turn (D7) and explores the following three possibilities. First, the property concerned may be identical with one already entered in SUBCSRIPT (D3); if so, the entry again remains unchanged. Second, it may be incompatible with one of the existing entries (D4): a property is *incompatible* with another property, we will say, if both are members of the same category. If so, the property referred to by the base component is substituted for the entry concerned (D6). Finally, it may be neither identical nor incompatible with any of the properties entered; in that case, it is simply added as a further entry (D5). (A more elaborate routine might delete from SUBSCRIPT any entry *x,* such that no word could have the property *x* and, in addition, have the further property just entered. But this is not strictly necessary.) To illustrate, suppose that SUBSCRIPT and INDEX are as above; the first rule, as we remarked, will be rule 17. The base component of this rule refers to a property singular which is identical with none of the initial entries but which is incompatible (since it too is assigned to the category NUMBER) with the entry

plural. By D6, SUBSCRIPT accordingly will be altered to read:

SUBSCRIPT Pa
Fu
Ind
3
sg

2. The index symbol in the base component is substituted for the existing entry in INDEX. In the case of rule 17, INDEX would of course again read

INDEX V.

On the next pass, however, the rule identified by Section B would be rule 10; at that point, INDEX would accordingly be altered to read

INDEX S ,

SUBSCRIPT, on this pass, remaining unchanged. In this way, the base component of each succeeding rule determines the conditions which the reference component of the next rule will have to satisfy; the cycling ends (see C 2, above) only when a rule is found with *R* as its base component. When it does end, the operations accumulated in OPERATION STORE supply the realization of the word which determined the initial entries.

IV. Discussion

The strategy discussed in Sections II and III may be profitably compared with the lexeme-to-morpheme encoding procedure suggested by Lamb (1964). Our two proposals have their inspiration in entirely different models of grammatical description; consequently, a decision between them should ideally be a matter of linguistic argument. Matthews (1965) suggests that each model is appropriate to a certain type of language. Lamb, on the other hand, appears to take it for granted that his model is appropriate to all. From the purely practical point of view, there seems to be three points that may be of importance.

1. A likely objection to the proposals put forward in Sections II and III is that the inflectional rules are ordered. This necessitates a separate pass through the grammar, or at best a pass through all rules whose reference components share the relevant index symbol, for each successive rule. To the majority of linguists, ordering should scarcely require justification. It has always been the practice to secure a generalization (e.g., those expressed by rule 3 or rule 31) by allowing any such generalization to have stated exceptions (e.g., those expressed by 1-2 or 24-30); in interpreting a grammar such exceptions must clearly be considered before the general rule becomes eligible to be applied. But, of course, this practice is not strictly necessary. An unordered set of rules will merely tend to be longer than its ordered equivalent. In any application, one must therefore choose what seems to be the lesser of two evils: either one must enlarge the grammar (to achieve what may be a speedier lookup), or one must tolerate a more tedious procedure (to achieve a more compact grammar).

2. An equally nugatory objection concerns the introduction of morphological operations. This approach appears to be justified on linguistic grounds. Numerous examples of "replacive morphs" (e.g., the replacement of the stem nucleus by *a* in English *sang, ran,* etc.) attest the advantages of a "process" as opposed to an "arrangement" model of morphological description. But the associated routine is more cumbersome. Applying the operations must form a separate part of the encoding procedure; furthermore we have introduced at least one operation (symbolized by SFV in rules 33 and 34) which is of an awkwardly sophisticated kind. However, it is possible to write a grammar that would be equivalent to the one in Section II but that would refer to suffixes instead of operations; it would merely be longer and would obscure, to the eyes of this linguist at least, the nature of the moveable accent. Similarly, it is possible to concoct an "arrangement" solution for the strong verbs in English, for example, by enlarging the inventory of morphophonemes and associated phonological rules. Again, therefore, one has to strike a balance. Either one must make what may be a real sacrifice in descriptive elegance, or one must put up with the more tiresome procedure.

3. There is at least one more serious criticism; namely, that we have ignored the problems of compounding and of "derivational" (as opposed to inflectional) morphology. According to the accepted morphemic model, the *con* in *condurrébbe* or the *s* in *slacciare* are handled no differently from the *ebb, ar,* etc.: there are morphemes, say *{con}* and *{s},* which have allomorphs *con* and *s* in the same way that other morphemes, say {Future}, {Infinitive}, etc., have allomorphs *r, ar,* and so forth. How would this work out in terms of the model in Section I? There are, of course, two trivial answers to this question. The first is to treat the compounding or derivational element as a further morphosyntactic property. For example, one might assign to *condurrébbe* the syntactic representation

$$\text{DURRE}_{\text{con, Fu, Pa, Ind, 3, sg}}$$

(using a fake Infinitive to symbolize the lexeme); its realization might then be handled by substituting *X* for *R* in rules 9, 23, etc., and adding, *inter alia,* a rule:

$[\text{X}_{\text{con}}]$ Prefix con, *R*

Alternatively, one could say that all compound and derived lexemes require a separate dictionary entry:

the prefix *s* would simply be part of the root of SLACCIARE, the *con* part of the root of CONDURRE, and so forth. Neither, however, would represent more than a trivial solution. It is unattractive to list all such lexemes in the dictionary, since some have a meaning (e.g., a translation meaning) which may be predicted from the entries for the separate elements. On the other hand, it is notorious that this is not always the case: why, therefore, should these elements receive the same treatment as semantically regular morphosyntactic properties? The problem of derivational morphology is a serious problem, for which no one (to my knowledge) has yet proposed a satisfactory solution.

Received December 10, 1965

References

Hall, R. A. *Descriptive Italian Grammar.* (Cornell Romance Studies, Vol. 2.) Ithaca, N. Y.: Cornell University Press, 1949.

Lamb, S. M. "On Alternation, Transformation, Realization, and Stratification," *Monograph Series on Languages and Linguistics,* Vol. 17 (1964), pp. 105-22.

Matthews, P. H. "The Inflectional Component of a Word-and-Paradigm Grammar," *Journal of Linguistics,* Vol. 1 (1965), pp. 139-71.

Reynolds, B. *Cambridge Italian Dictionary,* Vol. 1: *Italian-English.* Cambridge: Cambridge University Press, 1962.