Transactions of the Illinois Academy of Science
(1989). Volume 82, 1 and 2, pp. 39-46

ON THE LINGUISTIC SIGNIFICANCE OF THE METAGRAMMAR IN TWO-LEVEL MODELS OF NATURAL LANGUAGE

J.V. Rauff
Millikin University
Decatur, IL 62522

ABSTRACT

The metavariables of two-level grammars used to model natural language have generally served as *ad hoc* features needed to avoid transformations in parsing and generation. Here it is suggested that the metavariables can serve to encode the communicative function of a string in the language accepted by the grammar. This encoding can direct a mapping from a sentence in one natural language to its functional equivalent in another natural language. The process is exemplified in the context of automatic translation from Mandarin Chinese into English.

1. Introduction.

Two-level models of natural language have been under development since 1982 by Krulee (1985, 1987) and his students (Rager 1987, Larson 1987, Rauff 1988). Initially, the metagrammars of these two-level models served primarily as parsing devices, allowing rather problematic syntax, generally English, to be conveniently parsed by augmented transition networks. In this paper, I suggest that the metagrammar may have a linguistic counterpart. Recognition of this counterpart and its development could lead to two-level models that pay more attention to the semantics and pragmatics of natural language. More significantly, this linguistic counterpart is already receiving serious research attention by linguists involved in the teaching of foreign languages. The two-level approach to natural language processing then may be shown to offer a bridge between applied linguistics and computational linguistics.

I begin with a discussion of two-level grammars.

2. Two-level grammars.

Two-level grammars were first used in the development of the programming

language Algol 68 (see Van Wijngaarden 1976). A two-level grammar consists of two interacting formal grammars. In the protogrammar, certain constituents are independently derived by means of rules in the metagrammar. For example, consider the grammar G_1 (due to Krulee 1985):

metagrammar:

N ---> X | XN

protogrammar:

S ---> bn | bXS
X ---> a

The derivation of a string in G_1 begins with the S-production in the protogrammar. The presence of the **meta**variable N causes a shift of control to the metagrammar. When the derivation of a string in the metagrammar is completed, that string is substituted for N in the protogrammar, and the **proto**-derivation is continued. Once an assignment to a metavariable is completed, it may not be changed for the duration of the derivation. This requirement is called the **principle of uniform replacement.** A sample derivation is shown in detail below. The following notation is used:

===> indicates a one step derivation in the protogrammar.
===>> indicates a one step derivation in the metagrammar.
~~~ indicates a transfer of control from one grammar to the other.

Derivation of a string in $L(G_1)$:

S ===> bXS ~~~ N ===>> XN ===>> XXN ===>> XXX ~~~ bXXXS ===> baXXS ===> baaXS ===> baaaS ===> baaabN ~~~ N is XXX ~~~ baaabXXX ===> baaabaXX ===> baaabaaX ===> baaabaaa.

It is clear that $L(G_1) = \{ (ba^k)\ (ba^k)^* : k \geq 1 \}$.

The metagrammar enables a two-level grammar to generate rather complex languages with relatively simple rules. Indeed, it is possible to write two-level grammars with context-free components which generate some context-sensitive languages (Greibach (1974) has investigated the generative power of two-level grammars under various restrictions on the meta- and proto-grammars).

For example, the well-known context-sensitive language $L_1 = \{a^n b^n c^n : n \geq 1\}$ can be generated by the two-level grammar G_2.

G_2:

metagrammar:

X ---> IX | I

protogrammar:

S ---> A<X> B<X> C<X>
A<IK> ---> A<I>A<K>
B<IK> ---> B<I>B<K>
C<IK> ---> C<I>C<K>
A<I> ---> a
B<I> ---> b
C<I> ---> c

Here, the variables A<X>, B<X>, etc. bear the funny bracket extensions to denote reference to the metagrammar. Also, the recursive rules which begin with A<IK>, B<IK>, and C<IK> are actually shorthand for a large set of rules and potential rules. A sample derivation will illustrate how G_2 operates.
~~~

S===> A<X> B<X> C<X>~~~ X===>> IX===>> II~~~
A<II> B<II>C<II> ===> A<I>A<I>B<II>C<II>===> a A<I>B<II>C<II>
===> aaB<II>C<II> ===> aaB<I>B<I>C<II> ===>aabB<I>C<II> ===>
aabbC<II> ===> aabbC<I>C<I> ===> aabbcC<I> ===> aabbcc.

It should be clear that $L(G_2) = L_1$.

3. Two-level grammars and natural language processing.

Krulee (1985) has shown how the strategic use of two-level grammars in natural language processing can eliminate the need for transformations in a formal grammar of a natural language (English). In this respect, his work aims at the same goal as Gazdar's **Phrase Structure Grammar** (Gazdar and Pullum 1982), namely, the elimination of transformations. As an example of how Krulee can model natural language with two-level grammars, consider the grammar G_3 (adapted from Krulee 1985:31-32), which generates English yes/no questions.

G_3:

metagrammar:

Type ---> Decl | Ques

protogrammar:

S ---> X<Type> VP
VP ---> Verb NP2
NP2 ---> Det Noun
NP ---> Det Noun
X<Decl> ---> NP
X<Ques> ---> Aux NP

In this example, I have omitted the parts of Krulee's grammar which control number agreement and verb tenses. If we take these controls as given, then G_3 can generate sentences like

(1) The boy ate the carrots.
(2) Did the boy eat the carrots?

In both cases, the protogrammar relinquishes control to the metagrammar early in the derivation in order to determine the primary function (interrogative or declarative) of the sentence to be generated. The derivation of (2) is shown here:

S===>X<TYPE> VP~~~ TYPE===>> QUES~~~ X<Ques> VP ===> Aux NP VP
===> Did NP VP ===> Did Det Noun VP ===> Did the Noun VP ===> Did the boy
VP ===> Did the boy Verb NP2 ===> Did the boy eat NP2 ===> Did the boy eat Det
Noun ===> Did the boy eat the Noun ===> Did the boy eat the carrots?

Krulee writes that the metagrammar's "primary function is to select among declaratives vs. interrogatives . . ." (Krulee 1985:31). However, it is clear from his choice of names for the metavariables that he attaches no linguistic significance to these metavariables. Yet, the selection of "declarative vs. interrogative" is certainly a linguistic choice made in the utterance of the sentence.

This choice must be made in any natural language. It is therefore not unreasonable to expect to see certain similarities in the metavariables of two-level grammars of a variety of languages. The question which then arises concerns the number and nature of these common metavariables. That is, besides making statements and asking questions, what other functional choices are made in uttering a sentence in any natural language?

4. Functional-notionalism.

A preliminary answer to the question posed in the last paragraph has come from a surprising (for computational linguistics) source, foreign language teaching research.

In an influential work, Van Ek and Alexander (1980) specify the learning objectives for students of a foreign language. These objectives not only enumerate what is needed in the learning of a foreign language, but also suggest some fundamental aspects of language when it is viewed as a communicative skill.

Van Ek and Alexander's objectives (1980:7-9) address the following components:

1. the situation in which the language is to be used,
2. the language activities in which the learner will participate,
3. the language **functions** which the learner will fulfill,
4. the general and specific language **notions** which the learner will be able to handle,
5. the language **forms** which the learner will be able to use, and
6. the degree of skill with which the learner performs these language tasks.

As an example of these components consider the language skills a student needs for asking a question in an algebra class. The situational requirements include the topic at hand, let's say finding the roots of polynomials, the respective roles of student and teacher, and the degree of formality in the classroom. The student's activities include formulating and asking the question, determining whether or not the question has been answered, and understanding the response. The language function to be performed is that of asking a question. The notions involved could include sequence ("Do we always look for the rational roots first?), or temporality (Could you explain why 6 is not a root, **again**?) and specific notions like **root, polynomial, factor, coefficient**, etc. The language forms encompass the proper syntax for the desired function. In this situation, the forms include sentence initial auxiliary verb constructions (Do we . . .?), Question-word constructions (Why . . .?), etc. Finally, the level of skill with which the question is posed is probably proportional to the effectiveness of the question in eliciting a useful response. (Contrast "I don't get why you do that to those things?" To "Why did you divide every coefficient by 2?")

Viewing language learning as learning the appropriate forms and notions to carry out particular functions in specified situations has gained wide popularity among teachers of foreign languages and teachers of English as a second language (Finocchiaro and Brumfit 1979) and has been adopted as a paradigm for ESL textbooks (Zaffran, et al. 1988). For computational linguistics, this **functional-notional** approach to language learning provides a mechanism for constructing a formalism that recognizes language as communication.

5. Functional-notionalism and two-level grammars.

In combining the pedagogy of the functional-notional approach with the formalism of two-level grammars, we are implicitly adopting the idea that the form of language is in some way at least partially determined by the function it is intended to perform. The notion that linguistic form and linguistic function are interrelated is by no means new. Indeed, it has been a persistent topic of linguistic discussion (see Robins 1979, or Dinneen 1967). Most recently, the most detailed

investigation of the "complex interaction of form and function in language" has been undertaken by Foley and Van Valin (1984), Silverstein (1981), and Halliday (1979). It is not the aim of this paper to get involved too deeply in the linguistic arguments revolving about form and function. Rather, I am concerned with the formal generative implications of the discussion.

If we acknowledge that, at least some linguistic forms follow from the linguistic function they are intended to perform, then it is not a great leap to postulate that perhaps a generative grammar of a language could be constructed at two levels. A **meta**grammar would generate the specific functions which a string is intended to perform, and then these functions would direct the operation of a **proto**grammar which would generate the specific linguistic entity required to perform those functions.

The two-level model (G_3) of the small fragment of English shown in section 3 adopts this premise. A larger two-level grammar built along the same principles would follow a careful analysis of form and function in a specific natural language. Although a complete grammar along these lines does not exist for any language (see Langendoen and Postal (1986) for an interesting argument denying that such a grammar is even theoretically possible), we can find partial analyses in language texts that adopt some version of functional-notionalism. In these texts a student is given the appropriate language **forms** needed to perform certain language **functions.**

As an example of this consider DeFrancis' *Beginning Chinese* (1976). DeFrancis presents numerous "patterns" which are appropriate for fulfilling different functions. These include, asking a yes/no question, talking about the existence of something at a place, asking for and giving directional information, expressing purpose, etc. Rauff (1988) has taken 24 of these linguistic functions and their Chinese forms and constructed a two-level grammar based upon the principles outlined in this paper. Rauff's two-level grammar of Chinese (Rauff 1988:130-138) can generate the Chinese sentences given in the first 9 lessons of DeFrancis' text.

6. Application to machine translation.

One rather obvious application of the two-level functional-notional approach to modeling natural language is automatic translation between natural languages. Under the approach outlined in this paper, a translation of a sentence S_L in language L to an equivalent sentence S_M in language M would involve the following steps:

T1: Parse S_L and record the function and the specific notions.

T2: Generate the appropriate general form and specific notions in S_M necessary to perform the function determined in T1.

If we accept that an "equivalent" sentence exists in the target language (see Katz 1981), then the problem becomes one of determining the function of a string from its form. This is by no means a trivial problem! Questions of pragmatics enter into the discussion rather quickly as we try to deduce the intended function of "Can you reach the salt?" from its form. A great deal of linguistic research has been expended on the relationship between language form and language function (see Foley and Van Valin 1984, and Levinson 1984). However, in many cases the communicative function of a sentence is retrievable from its syntax. In these cases the ability to determine the function may be computationally exploited.

As an example, the appendix gives some sample output of the translating program FTL (Rauff 1988) which uses the translating scheme described above and

two-level functional grammars of Chinese and English. (The numeral suffixes on the Chinese lexical items denote tones.) The Chinese sentence "Talmen zai4 zher4 mai4 shen2mo", for example, is parsed by the two-level Chinese grammar in FTL. The parsing determines that the function of the sentence is a **WHAT-question,** and the notions include selling objects in a deictically referenced place. The English two-level generator then takes these functions and notions and produces the English string "What do they sell here?".

The basic process of the translating program is as follows. The two level-parser operates on an elementary augmented transition network that determines the function of the sentence from the syntax. The functions are derived from the Chinese "patterns" given in DeFrancis' (1976) beginning Chinese text. After the parse, the Chinese lexical items are mapped into English using an annotated Chinese-English dictionary. These English items are then passed, along with the sentential function, to an English two-level generator (also based on an augmented transition network). The English sentence is then produced according to the English grammar directed by the appropriate assignments of function in the metagrammar.

The use of communicative function is not the only approach to machine translation possible. Other systems are totally syntactic or rely on some sort of semantic representation of the input string (see Slocum (1985) for a survey of machine translation). However, the use of a metagrammar in a two-level grammar to model the communicative function of a sentence offers a new, perhaps more natural approach, to automatic translation.

7. Summary.

I have argued that the metagrammar of a two-level model of natural language may be used to control the linguistic function of a string being generated by the grammar. This type of modeling requires that some definite relationships between linguistic form and linguistic function be established. These relationships are most clearly spelled out in the pedagogy of functional-notionalism. One application of this approach to natural language processing is that of functional machine translation.

LITERATURE CITED

DeFrancis, J. 1976. Beginning Chinese. Yale University Press, New Haven.

Dineen, F.P. 1967. An Introduction to General Linguistics. Georgetown University Press, Washington, D.C.

Finocchiaro, M., and C. Brumfit. 1979. The Functional-Notional Approach. Oxford University Press, Oxford.

Foley, W.A., and R.D. Van Valin. 1984. Functional Syntax and Universal Grammar. Cambridge University Press, Cambridge.

Griebach, S.A. 1974. Some restrictions on W-grammars. Proceedings of the Annual ACM Symposium on the Theory of Computing: 256-65.

Gazdar, G., and G.K. Pullum. 1982. Generalized Phrase Structure Grammar. Indiana University Linguistics Club, Bloomington.

Halliday, M.A.K. 1979. Explorations in the Function of Language. North-Holland, Amsterdam.

Katz, J. 1981. Language and Other Abstract Objects. Blackwell, Oxford.

Krulee, G.K. 1987. Two-Level Processing of Natural Language. Indiana University Linguistics Club, Bloomington.

Krulee, G.K. 1985. Two-Level Representations of Natural Language. Indiana University Linguistics Club, Bloomington.

Larson, T.J. 1987. Semantics for coordinated substitution grammars as implemented in Prolog. Ph.D. dissertation, Northwestern University.
Levinson, S.C. 1984. Pragmatics. Cambridge University Press, Cambridge.
Rager, J.E. III. 1987. Multi-level Structures for Natural Language Processing. Ph.D. dissertation, Northwestern University.
Rauff, J.V. 1988. Machine Translation with Two-Level Grammars. Ph.D. dissertation, Northwestern University.
Robins, R.H. 1979. A Short History of Linguistics. Longman, London.
Silverstein, M. 1981. Case marking and the nature of language. Australian Journal of Linguistics 1:227-246.
Slocum, J. 1985. A survey of machine translation: its history, current status, and future prospects. Computational Linguistics 11:1-17.
Van Ek, J.A., and L.G. Alexander. 1980. Threshold Level English. Pergamon Press, Oxford.
Van Wijngaarden, A. et. al. (Eds). 1976. Revised Report of the Algorithmic Language ALGOL 68. Springer-Verlag, Berlin.
Zaffran, B., D. Krulik, and M. Scheraga. 1988. Hello, English. National Textbook Company, Lincolnwood, IL.

SAMPLE OUTPUT OF FTL

This appendix contains several examples of the output produced by the functional two-level translator, FTL. For each sample, the Chinese input sentence is given with the translation produced by FTL immediately below. The parentheses are present because the input and output sentences are both LISP expressions.

(WO3 BU HUI4 SHUO1 ZHONG1GUO-HUA4)
(I CAN NOT SPEAK CHINESE)

(CHENG2 LI3 TOU DE SHAN1 DOU1 HEN3 XIA03)
(THE MOUNTAINS INSIDE THE CITY ARE ALL VERY SMALL)

(TU2SHU1GUAN3 JIN1TIAN MEI2 YOU3 REN2 KAN4SHU1)
(THERE ARE NOT ANY PEOPLE READING IN THE LIBRARY TODAY)

(GAO1 TAI4TAI YOU3 PENG2YOU CONG2 MEI3GUO DAO4 ZHONG1-GUO LAI2 MA)
(DOES MRS GAO HAVE A FRIEND COMING FROM AMERICA TO CHINA?)

(TA1MEN AZI4 ZHER4 MAI4 SHEN2MO)
(WHAT DO THEY SELL HERE?)

(WO3 BU ZHIIDAO4 SHI ZAI4 JIA1 NIAN4SHU1 HAO3 HAI2SHI ZAI4 TU2SHU1GUAN3 NIAN4SHU1 HAO3)
(I DO NOT KNOW IF IT IS BETTER TO STUDY AT HOME OR STUDY AT THE LIBRARY)

(WO3 BU XIANG3 ZAI4 NER4 MAI3 SHU1)
(I DO NOT PLAN TO BUY A BOOK THERE)

(TA1 BU XIANG3 ZAI4 SHU1DIAN4 MAI3 SHU1)
(HE DOES NOT PLAN TO BUY A BOOK AT THE BOOKSTORE)

(WO3 MING2TIAN QU4 MAI3 MAO2BI3)
(I WILL GO TO BUY A WRITING BRUSH TOMORROW)

(HUO4ZHE3 WO3 MING2TIAN QU4)
(PERHAPS I WILL GO TOMORROW)

(NI3 ZHAO3 SHEN2MO)
(WHAT DO YOU LOOK FOR?)

('ZUO3 BIAR1 DE FANG2ZI LI2 YOU4 BIAR1 DE FANG2ZI DUO2MO YUAN4)
(HOW FAR IS THE HOUSE ON THE LEFT FROM THE HOUSE ON THE RIGHT?)

(MAO2BI3 ZAI4 ZHER4 MA)
(IS THE WRITING BRUSH HERE?)

(TA1 DE FANG2ZI ZAI4 SHAN1 SHANG4)
(HIS HOUSE IS ON THE MOUNTAIN)

(ZHONG1GUO-FAN4 HAO3 CHI1)
(CHINESE FOOD IS GOOD TO EAT)

(ZHEI4 BEN3 SHU1 RONG2YI NIAN4)
(THIS BOOK IS EASY TO READ)

(NEI4 GE GONG1YUAN2 HEN3 DA4)
(THAT PARK IS VERY BIG)

(TA1 XI3HUAN NEI4 GE SHU1)
(HE LIKES THAT BOOK)

(WO3 DE FANG2ZI ZAI4 NEI4 GE SHAN1 SHANG4)
(MY HOUSE IS ON THAT MOUNTAIN)

(WO3 DE FANG2ZI ZAI4 TA1 DE SHAN1 SHANG4)
(MY HOUSE IS ON HIS MOUNTAIN)

(NI3 HUI4 BU HUI4 SHUO1 ZHONG1GUO-HUA4)
(CAN YOU SPEAK CHINESE?)

(TA1 SHI4 YING1GUO-REN2)
(HE IS ENGLISH)

(GAO1 TAI4TAI YE3 SHI4 YING1GUO-REN2 MA)
(IS MRS GAO ALSO ENGLISH?)

(TA1MEN DOU1 HUI4 SHUO1 ZHONG1GUO-HUA4)
(THEY ALL SPEAK CHINESE)