Syntactic Analysis Requirements of Machine Translation

by

S. R. Petrick

IBM T.J. Watson Research Center

SYNTACTIC ANALYSIS REQUIREMENTS OF MACHINE TRANSLATION

S. R. Petrick

In this note I will confine my attention to machine translation (MT) systems which are based upon an underlying formal generative grammar. This is not to deny the potential importance of various computational aids to human translation, nor to deny the possibility of machine translation not based on a formal grammar. It is clear, however, that for fully automated MT any attempt to make use of presently existing linguistic theory or of that which is likely to exist in the foreseeable future requires a grammar-based approach.

A second assumption I wish to make is the existence of two distinct components of a grammar - - a syntactic component and a semantic component. The former assigns structure to sentences and the latter interprets those structures by translating them to a natural language (in the case of MT) or to an artificial language which has its own computer interpreter. It will not be assumed that the syntactic and semantic components necessarily interact in a simplistic fashion, i. e. , every syntactic output is to have a distinct well formed semantic interpretation, and the final output of the syntactic component is the input to the semantic component. Instead, we will, for example, allow the syntactic component to generate structures which are rejected by the semantic component, and we will allow semantic analysis (and rejection) of fragments of a syntactic structure prior to the complete determination of that structure.

The importance of the syntactic component has been recognized for some time. For the purposes of MT it has two distinct ends to achieve: on the one hand it must specify a large enough subset of the source language to meet the operational requirements of the MT application in question. (The related function of ruling out syntactically ill-formed sentences is of limited importance in MT). On the other hand the structures it assigns must provide a reasonable basis for semantic interpretation. These two requirements are closely related, i. e. , it is relatively easy to satisfy one at the expense of the other, but much harder to adequately meet them both.

A not uncommon attitude which has been expressed both in the computational linguistic literature and orally at symposia and conferences is that syntax in general and syntactic analysis in particular has been well worked over, is thoroughly understood, and presents no serious problems — in contrast to the situation in semantics where little has been done and not much is understood. I submit that such remarks reflect the experience of one who has chosen a class of grammars, in most cases context-free grammars, which permits a reasonable coverage of a source language at the expense of assigning structural descriptions which bear little relationship to underlying meaning and which, therefore, provide an inadequate basis for semantic interpretation. It is not just because large-coverage context-free grammars have been found to often assign 100 or more structural descriptions to unambiguous sentences that makes them inadequate. Rather, this is just symptomatic of a more deep-seated inability to relate form to underlying meaning.

This shortcoming is not limited to the class of context-free grammars. If the rewriting system is extended to encompass context-sensitive grammars and/or rewriting rules with whose constituents complex features can be associated then economies and linguistic generalizations are realized, but the fundamental problem of relating form to meaning appears intractable for any system which attempts to interpret the surface form of sentences. It was this realization that prompted Chomsky to propose as the basis for semantic interpretation deep structures which were in many cases far removed from surface structures. Chomsky made use of a transformational component to relate corresponding deep and surface structures, but the acceptance of the deep-surface structure distinction is a matter which is independent of any consideration of the most appropriate means for making explicit that correspondence. Accordingly, a host of models (each of which is a proposed linguistic theory even if not called such) have been proposed for mapping surface structures into corresponding deep structures, or (in some cases) for directly assigning deep structure to sentences without explicitly producing surface structure.

It is my contention that linguistic models which do not provide the deep structure of sentences (at least implicitly if not explicitly) fail to provide a basis for the semantic analysis of all but a small class of sentences, a class so restricted that its use is precluded for most applications including MT. Hence, for the remainder of my discussion I will focus my attention on the problems of syntactic and semantic analysis associated with some type of deep structure model.

As pointed out previously, there is a trade-off possible between syntax and semantics. If more is done by the syntactic component the task of the semantic component is lightened and vice-versa. Contemporary linguistic theory has been much concerned with this question of where to draw the line, and even though the questions of overall simplicity considered have not been motivated by any concern for MT, it is nevertheless instructive to consider the applicability to MT of models of present-day deep structure complexity. There is, of course, no general agreement among linguists as to the type and complexity of deep structures and of the related transformational component required by those deep structures. Even though these decisions loom large and important to linguists, however, they are not so large as to preclude assessment of the suitability of a rather large class of deep structure models for MT.

Let us begin then by considering the requirements of the semantic component. It is, of course, possible to produce sentences whose semantic analysis and/or translation requires not only a number of deep structure distinctions but also a large amount of information about the world, about logical deduction, and about the context of discourse in which the sentence appears. I am resigned to the prospect that these obstacles preclude for the foreseeable future extremely high quality translation.

My own experience with semantic interpretation has been with translation to a formal language which, although not a programming language in the sense of having an existing hardware or software interpreter, is close enough to a programming language that the task of translating it to an existing programming language is an easy one. The problem of translating a given structure to a functional programming language appears to me to be greater than that of translating that structure to another natural language. This follows from two considerations: First, deep structures of different languages which have been proposed to date are remarkably similar. In those cases where differences have been argued, they have seldom exceeded differences in subject, verb, object ordering. Deep structures differing so slightly are easily related through the use of such standard translation mechanisms as the Irons Translator[1]. .

Second, the task of using a transformational grammar to convert deep structures into surface structures is not conceptually difficult. Hence, it would appear that for a very large class of sentences, the translation sequence shown below should provide the basis for translation:

$$\begin{tikzcd}
\begin{array}{c}\text{Source}\\ \text{Sentence}\\ \text{S}\end{array} \arrow[r] & \begin{array}{c}\text{Deep Structure}\\ \text{of S}\end{array} \arrow[r] & \begin{array}{c}\text{Corresponding}\\ \text{Target Language}\\ \text{Deep Structure}\end{array} \arrow[r] & \begin{array}{c}\text{Target}\\ \text{Language}\\ \text{Translation}\\ \text{of S}\end{array}
\end{tikzcd}$$

Indeed, it has been my experience that semantic interpretation of deep structures through the use of the Irons or more generally the Knuth[2] translation mechanism even provides a reasonable basis for a natural language question answering system. This has also been argued by Kellogg and Thompson among others. For more discussion see reference 3.

I have argued that the use of a deep structure (read semantic structure if you prefer) generative grammar does provide a reasonable basis for MT. It does so, however, by throwing a considerable burden on the syntactic component. We have seen that structures can be assigned which appear adequate for the purposes of MT. But what of the coverage requirement, i. e., that a sufficiently large subset of the source language be specified? In addition, we must concern ourselves with the theoretical and practical requirements of syntactic analysis for a class of grammars that is capable of assigning adequate deep structures.

I will discuss these two considerations with respect to generative transformational theory and also, more briefly, with respect to other deep structure-based linguistic theories.

Let us first consider the matter of coverage. It is, of course, the case that most transformational studies of syntax do not supply completely specified base and transformational component rules in discussing syntactic phenomena. There have been, however, a few attempts to write a completely specified set of rules within a well-defined transformational framework [4, 5, 6, 7, 8]. These efforts establish a lower bound on coverage which can be achieved without sacrificing structural adequacy. It is somewhat difficult to characterize the coverage achieved by any means short of exhibiting the grammars in question. There are, however, at least two ways to give a feel for the coverage attained by a specific grammar. The first is to give a list of "representative" sentences and the second is to list the syntactic constructions and phenomena provided for. Thus, for example, Rosenbaum[7] gives derivations of the 22 sentences:

1. the boys like the girl

...

21. the pajamas of a king are colorful
22. the people who approve of him think that John is smart

He also lists 79 representative sentence types and includes transformations for handling verb phrase complements, pronominalization, preposition segmentalization and raising, indirect objects, relatives, genitives, negatives, certain time and place adverbials, etc. Similarly, in a more recent effort at the IBM Thomas J. Watson Research Laboratory a transformational grammar has been produced which generates such sentences as:

> _what companies had a profit which was more than ten million dollars_?

and _print the one element of the set which contains M which is atomic_ and provides such construction types as: yes-no and Wh-questions, passives, prepositional phrases, nominal structures formed from underlying abstract verbs, restrictive relatives, possessive genitives, and certain types of negatives, comparatives, and coordinate structures.

Now just as existing grammars establish a lower bound on coverage attainable there are several considerations which suggest upper bounds for at least the foreseeable future. For example, many syntactic phenomena may be identified which have not yet been studied by anyone. Many other phenomena have been studied, but the results have served more to show the existence of substantial problems than to offer compelling and widely accepted solutions. Examples here are plentiful and include coordination, gapping, and pronominalization as well as almost every syntactic phenomenon which has been studied to some extent. And finally, experimental work conducted to date shows that it is far from trivial to put together and test grammars that provide for such relatively well understood constructions as yes-no questions, WH-questions, restrictive relatives, imperatives, etc.

The large number of unexplored and little understood syntactic phenomena suggest difficulty in achieving sufficient coverage for practical application, but an even more instructive exercise in illustrating this difficulty is provided by producing a set of sentences thought to be useful and representative for some application and comparing their syntactic requisites with the facilities offered by any existing or proposed grammar. I have seen this operation carried out at the MITRE Corporation with respect to a command and control question answering application and have myself undertaken the same task for a formatted file question answering facility. The results were the same. Very low coverage was observed; certainly less than 10% of the sentences studied were covered even allowing for lexical addition and extension by including some rather obvious additional transformations. The saving feature in the case of natural language question answering systems or natural language programming systems, however, is that they need not process unconstrained input sentences. Instead the user can be constrained to and instructed as to how to limit his input in terms of both lexicon and allowable constructions. All that is required is that natural subsets provided must be learnable by human speakers and must be rich enough to permit expressing that which must be expressed in a convenient fashion. The attainability of even these requirements remains to be established but at least offers some hope of success. On the other hand the usual situation with MT is that the input is not produced with the limitations of a particular formal grammar in mind. This, more than any other single factor, convinces me that grammar-based MT offers little hope for practical usage for at least the next ten years. This is not to say that MT is not an interesting and productive vehicle for keeping linguistic research in both syntax and semantics tied to reality. Others might disagree with this assessment, of course.

There may be a few MT applications where time and economic considerations permit the phrasing or rephrasing of source sentences by speakers cognizant of a system's grammatical constraints. Such an example is the preparation of technical manuals in one language for translation into another language. This is, however, not the usual situation in MT.

When we leave the (at least for me) familiar grounds of transformational theory and consider the coverage problem for such analysis-based linguistic theories as those of Woods[9], Winograd[10], Bobrow and Fraser[11], Thorne,[12] Moyne,[13] Kellogg,[14] Kay[15] and Simmons,[16] we are faced with a difficult task for a number of reasons. Many of these models have been used only sparingly for the specification of any natural language. Hence, there is little to go on in assessing the coverage of these models. In addition, those models for which one or more large grammars have been written have not been documented in a way and to an extent which makes the determination of coverage feasible. Alternative clarification of coverage via sample sentences and listed construction types presents the same problem as we observed for transformational grammars, but whereas most linguists are by this time familiar with transformational formalism, this is not true of the aforementioned analysis-based models. Therefore, their coverage can at present be estimated only by their originators. It is far from clear to this observer that these approaches offer the same independence of construction types as is achieved by transformational theory. In any case, none of these models have supported claims of greater coverage than that afforded by current transformational theory. It is important to note that although these models are often described as "transformational" by their originators, they have not been related to transformational theory and hence must be judged on the usual grounds of linguistic adequacy just like any other proposed linguistic theory.

The remaining consideration is the theoretical and practical requirements of syntactic analysis for a deep structure - specifying class of grammars. For those analysis-based grammars previously mentioned there are few theoretical syntactic analysis problems. In addition, the computation time required for parsing, although generally not known, could reasonably be expected to be less than that required for parsing with respect to a transformational grammar. (Whether it is sufficiently small to satisfy economic considerations is, of course, another story.) This is to be expected for analysis-based linguistic theories whose principal motivation is to facilitate syntactic analysis. It is descriptive adequacy, not syntactic analysis considerations which are most likely to preclude the practical use of analysis-based grammars.

The situation is quite different with respect to transformational grammars. There is no shortage of work in linguistic description through the use of transformational grammars, although it must be noted that most efforts are directed toward determining the allowable class of transformational grammars rather than toward developing in detail any one comprehensive grammar. Syntactic analysis for any class of transformational grammars is a very complex and time-consuming proposition. It is probably for this reason that most workers in computational linguistics have chosen to forego conventional transformational theory in favor of an analysis-based alternative.

There have been only two computer implemented efforts on transformational grammar syntactic analysis. One, carried out by the MITRE Corporation, was limited to a particular grammar; a syntactic analysis program was tailored to this grammar. The program appeared to be successful in producing desired structures in a reasonable time, but it was never established that this program invariably found all of the structures assigned to a sentence by the particular transformational grammar in question (i. e. , that it was, in fact, an analysis program for that grammar).

In contrast to the MITRE approach, Petrick[17] defined a class of transformational grammars and found a syntactic analysis algorithm that is valid for members of this class. The extremely nondeterministic nature of this algorithm made unfeasible the treatment of grammars as written by a linguist unfamiliar with the analysis procedure. However, Kirk and Keyser[6] showed that by suitable recasting, a substantial portion of an existing grammar (due to Rosenbaum) could be used for syntactic analysis.

In addition to the problem of computing time, there is another serious difficulty in transformational grammar syntactic analysis. The class of grammars for which syntactic analysis algorithms have been devised does not include many of the facilities currently being used by descriptive grammarians. Indeed, transformational theory is far from static, and at any given time there is little agreement on just what should constitute an allowable class of transformational grammars. In reference 18 we give an account of the current status of syntactic analysis for transformational grammars. In summary, it can be stated that although the class of grammars for which syntactic analysis is possible has been significantly extended, the introduction of new variants of transformational theory has more than kept pace with theoretical and programming efforts to cope with them. Consequently, any given linguist would undoubtedly find that his rules and assumptions do not correspond perfectly with the formulation of the allowable class of grammars. Nevertheless, it is hoped that this class is now extensive enough to permit recasting of current transformational grammars into an acceptable form without seriously compromising their linguistic integrity.

REFERENCES

1. Irons, E. T. A syntax directed compiler for ALGOL 60. *Comm ACM 4* (Jan. 1961), pp. 51 - 55.

2. Knuth, D. E. Semantics of context-free languages, *Math. Sys. Theory 2* (1968), pp. 127 - 145.

3. Petrick, S. R. On the use of syntax-based translators for symbolic and algebraic manipulation, *Proc. Second Symp. on Symbolic and Algebraic Manipulation*, Los Angeles, Calif., March 1971, pp. 224 - 237 (Also IBM RC3265)

4. Zwicky, A., Friedman, J., Hall, B., and Walker, D. The MITRE syntactic analysis procedure for transformational grammar *Proc. Fall Joint Computer Conference*, 1965, Spartan Books, Washington, D. C., pp. 317 - 326.

5. Rosenbaum, P.S. and Lochak, D. The IBM core grammar of English, *Specification and Utilization of a Transformational Grammar*, *Scientific Report No. 1*, (IBM Corp., Yorktown Heights, N. Y., 1966)

6. Keyser, S. J. and Kirk, R., Machine recognition of transformational grammars of English. Air Force Cambridge Res. Labs, final report No. 67-0316, Jan. 1967.

7. Rosenbaum, P. S., IBM English grammar II, *Specification and Utilization of a Transformational Grammar, Scientific Report No. 2*, (IBM Corp., Yorktown Heights, N. Y., Oct. 1967)

8. Stockwell, R. P., Schachter, P., and Partee, B. H. Integration of transformational theories of English syntax, USAF Electronic Systems Division Report ESD-TR-68-419, Oct. 1968, Vols. I, II.

9. Woods, W. A. Transition network grammars for natural language analysis, Comm. ACM 13 (Oct. 1970), pp. 591 - 606.

10. Winograd, T. Procedures as a representation for data in a computer program for understanding natural language, Rept. AI-TRI, Artificial Intelligence Laboratory, MIT, 1971.

11. Bobrow, D. G. and Fraser, J. B. An augmented state transition network analysis procedure, Proc. Internat. Joint Conf. on Artificial Intelligence. Washington, D. C. , 1969, pp. 557 - 567.

12. Thorns, J. , Bratley, P. , and Dewar, H. The syntactic analysis of English by machine, Machine Intelligence 3, D. Michie (Ed.), American Elsevier, New York, 1968.

13. Moyne, J. A., Loveman, D. B. and Tobey, R. G. , Cue: A preprocessor system for restricted, natural English, Proc. Symp. on Information Storage and Retrieval. Univ. of Maryland, April 1971, pp. 47 - 60.

14. Kellogg, C. , Burger, J. , Diller, T. , and Fogt, K. The Converse natural language data management system: Current status and plans, Proc. Symp. on Information Storage and Retrieval. Univ. of Maryland, April 1971, pp. 33 - 46.

15. Kay, M. , Experiments with a powerful parser, Proc. Deuxieme Conference International sur le Traitement Automatique des Langues, Grenoble, Aug. 1967, Paper .No. 10.

16. Simmons, R. F. , Burger, J. F. , and Long, R. E. , An approach toward answering English questions from text, Proc. 1966 Fall Joint Computer Conf. , 1966, pp. 357 - 363.

17. Petrick, S. R. , A recognition procedure for transformational grammars, Ph. D. thesis, MIT, 1965.

18. Petrick, S. R., Syntactic analysis for transformational grammars, Proc. of the Conference on Linguistics, The University of Iowa, Iowa City, Iowa, Oct. 1970.