Combining Rationalist and Empiricist Approaches to Machine Translation

Ralph Grishman
Computer Science Department
New York University
New York, NY 10003, U.S.A.

and

Michiko Kosaka
Computer Science Department
Monmouth College
West Long Branch, NJ 07764, U.S.A.

Abstract

Better methods are needed for acquiring the knowledge which must go into machine translation systems. The call for papers for this conference contrast two approaches: the rationalist (based on linguistic theory) and the empiricist (based on analysis of large corpora). We suggest in this paper an intermediate approach which draws on the strengths of both.

In this approach, parallel corpora in the source and target languages would be analyzed to produce parses and syntactically regularized tree structures. The individual source and target language trees would then be aligned, yielding a set of correspondences between source and target structures involving specific words. Classes of closely related words would be identified from a distributional analysis of the parsed corpora, and these classes would be used in turn to generalize the correspondences. These generalized correspondences would then serve as the transfer rules of a machine translation system.

1. Introduction

The call for papers for this conference contrasts two very different approaches to the development of machine translation systems, dubbed the rationalist and empiricist approaches. Although these are depicted as competing approaches, we believe that they are complementary, and that the appropriate question is how to develop a system which combines the two approaches in a way which takes advantage of the strengths of each.

Typical modern 'rationalist' systems involve several layers of sentence analysis and structural regularization, followed by transfer and generation ('deep transfer'), or by analysis and generation all the way to an interlingua. All of these layers must be carefully crafted by hand. Computational linguists have been relatively successful in crafting rules for the first levels of analysis,

which develop a phrase structure and map it into some functional structure. Work in linguistics and computational linguistics has given us tools (grammars) which have broad coverage within individual languages and even some commonality across languages. On the other hand, at the deeper levels of analysis we find that the system developer has to deal individually with a large number of language-specific linguistic and lexical idiosyncrasies. The difficulty in doing so has limited the ability to improve machine translation systems and port them to new domains.

Empiricist methods, in contrast, have tried to learn correspondences from large samples of paired bilingual corpora. The correspondences have typically been established at a shallow level. In the work on statistical translation at IBM Yorktown Heights (Brown et al. 1990)', for example, correspondences are established directly between word sequences. In the work on example-based machine translation at Kyoto (Nagao 1987, Sato and Nagao 1990) and ATR (Sumita and lida 1991), correspondences involve word sequences or surface analyses. This use of correspondences at a shallow level of analysis is both a strength and a weakness. A strength, because the method is then not limited by deficiencies of the manually-developed deeper analysis methods; a weakness, because generalities which may exist at a deeper level (e.g., at functional structure) will be lost.

We believe, therefore, that in the long term the most powerful methodologies for developing machine translation systems will come to combine the strengths of both approaches. A rationalist approach should be employed for those levels of analysis for which well-developed, broad-coverage linguistic theories are available. Empiricist approaches should be used to acquire the information which is more lexically or domain specific, and for which there is (as yet) no broad theoretical base. From this perspective, we anticipate that future empiricist systems will develop from current systems in three directions:

(1) **Establish correspondences between regularized syntactic structures.** Establishing correspondences at a deeper level offers two advantages. First, the correspondences are simpler. At a deeper level many of the basic word-order differences between languages can be eliminated, so that the discovery process can focus on more lexically-specific differences. Second, correspondences which would have to be discovered independently for different syntactic structures at surface structure can be identified once at a deeper level of analysis (for example, a particular verb-particle combination that would have to be recognized separately in main, relative, and reduced relative clauses in English surface structure could be treated as a unified phenomenon with only a small degree of syntactic regularization).

(2) **Reduce correspondences to rules.** Example-based machine translation, in particular, suggests an approach where the bilingual data base is consulted directly for translation purposes. It should be possible, however, to extract the correspondences from this data base and then collapse common correspondences into transfer rules (i.e., form generalizations). This would have two virtues. First, it would reduce the amount of data which need be consulted during translation. Second, by comparing similar rules, the attempt at generalization

would provide feedback regarding the word classification required in order to distinguish patterns which must be translated differently.

(3) **Use automatically-generated word classifications.** Word classification plays two important roles. First, in conjunction with case frames, it can be used for selection during parsing of the source text. Second, these word classes can be used to generalize correspondences into transfer rules. At present most systems rely on relatively broad classes which are established manually as part of dictionary construction. However, by automatically examining large numbers of bilingual correspondences, it should be possible to identify classes which are suitable for reducing these to a smaller number of transfer rules.

In addition, we have discovered that the generalizations obtained through the use of these word classes have far reaching effect in that the word class information together with the syntactic information dictated the form of the surface form of the target language. Thus, we were able to control the quality of the translation stylistically in some instances (e.g. passive sentences rather than active sentences). This observation argues strongly in favor of the reduction to rules based on word classes in some environments.

2. System overview

At the core of our design would be a transfer-based machine translation system of relatively conventional structure: several levels of analysis, producing a functional structure akin to LFG f-structure; a set of transfer rules; and a component for generating target strings from the intermediate structure. The simplest transfer rules would specify a head (a word or class of words in the source language) and a set of source language operands, each consisting of a syntactic marker and a constraint on the values associated with that marker. The constraint would typically be a word class. The rule would also specify a target-language head and, for each operand, the corresponding syntactic marker in the target language. More complex rules would have more than one level of tree structure in the source or target language, and thus allow for structural transfer.

The syntactic rules for analysis and generation would be hand coded. In the long term, we would expect all the other information to be acquired automatically: semantic word classes and patterns which would be used to disambiguate the input text, and transfer rules (which would also reference the semantic word classes).

This acquisition procedure would make use of a large aligned bilingual corpus and would proceed in several stages. We would first process the source language and target language texts separately in order to produce semantic patterns and initial sets of semantic classes for the two languages. We would then make use of the correspondences between the two languages to develop the transfer rules and refine the word classes.

2.1. Acquisition of semantic patterns and classes

For the acquisition of semantic patterns and classes, we would expand upon an approach for which we have already executed some preliminary experiments in English (Hirschman et al. 1975, Hirschman 1986, Grishman et al. 1986, Grishman and Sterling 1992). We would take a substantial sample of text, parse it, and regularize the parse trees. Since this would be done without semantic patterns, this would produce many parse trees for each sentence.

Each parse would then be broken down into its constituent S and NP structures; each structure would then be further divided into a set of triples of the form

head - syntactic function - value

where "value" represents the head of the constituent bearing the specified functional relationship to the head of the structure. For example, the sentential structure for "John eats cheese",

(S eat (tense past) (subject (np John)) (object (np cheese)))

would yield the triples

eat - tense - past
eat - subject - John
eat - object - cheese

We would count these triples over the entire corpus. If a sentence generated N parse trees, a triple produced from such a parse tree would be weighted by 1/N.

Over a large sample, semantically correct triples are likely to be seen repeatedly, while incorrect triples (due to incorrect parses) will be more scattered. Therefore these raw counts will already give some indication of the dominant semantic relations for the domain (Grishman and Sterling 1992). These counts may be refined by an iterative training procedure analogous to the inside-outside training for Markov models (Baker 1979). We may use these counts to infer probabilities for each generalized triple, and then use these probabilities to compute a probability for each parse tree (for sentences with multiple parses). These probabilities may in turn be used to refine the counts, with a triple being weighted now by the normalized probability of the parse tree, rather than by 1/N. The result of this process is both a set of refined counts for triples and probability assignments for trees; further processing may be limited to the most probable tree for each sentence.

These triples counts can then be used to compute similarity coefficients between words based on the number of common contexts in which they appear, and then to group into semantic classes the words with the highest similarity (Hirschman et al. 1975, Hirschman 1986). The semantic classes could then be used in turn to create generalized triples, referring in some cases to word classes instead of words.

2.2. The acquisition of transfer rules

After the semantic patterns and initial classes have been acquired, we will have two parallel corpora of regularized syntactic structures, with at least some degree of semantic filtering. The

transfer rules will then be acquired by a process of aligning the two sequences of trees.

For the purposes of tree alignment, we will consider a tree where the lexical head of a phrase dominates the head of all its arguments and modifiers. Thus, for the sentence "John eats cheese." we would have "eat" dominating "John" and "cheese". This is similar in appearance to a dependency structure, although in our case it is based on a regularized structure with syntactically labeled cases. We consider a lexical item which has no modifiers or arguments a terminal node, and any other item a non-terminal node.

The alignment process will be 'primed' with *a* bilingual dictionary, Based on these initial word correspondences, the procedure attempts to establish a correspondence between the nodes of the Japanese and English trees, operating bottom-up.[1] We will refer to the nodes of the source language tree as S_1, S_2, ... and the nodes of the target language tree as T_1, T_2,… We accept a correspondence between two terminal nodes if the bilingual dictionary licenses a correspondence between the lexical items (i.e., one of the translations of one of the lexical items is the other lexical item). We accept a correspondence between two non-terminal nodes S_i and T_i, representing source and target phrases, if, for every node S_j which is dominated by S_i and for which a corresponding node of the target tree, T_j, has been identified, T_i dominates T_j.

These constraints by themselves would permit a large number of different possible matches between two trees, where each match is a set of [S_i, T_i] pairs. We therefore score the matches based on several criteria and select the highest-scoring match. The criteria currently include:

- the number of node correspondences
- for each correspondence between non-terminal nodes, whether the bilingual dictionary licenses a correspondence between the lexical items
- for each correspondence between non-terminal nodes, the distance to the nodes below that node which are already paired up

This structure-matching process can also be used to resolve some syntactic ambiguities in the source or target language. If there are several high-probability parse trees on either side, we can attempt this structural match with each such tree and select the tree which yields the highest-scoring match.

Certain types of structural mismatch can be anticipated. For example, a prepositional phrase in English will frequently correspond to a full clause in Japanese. The matching process will need to take these into account by imposing a lesser penalty for such a correspondence than for other structural mismatches. Such a case is included in the examples we present below.

Once the trees have been (partially) aligned, they can be segmented at the corresponding nodes. The result will be a set of corresponding subtrees. In cases where the two original trees

[1] The matching would be done bottom-up because our manual studies showed excellent correspondence at the lower levels of the tree, with greater structural differences near the root.

are isomorphic, each subtree may be just a single level of the original tree. Quite frequently, however, we will have a single level in one tree correspond to several levels in the other tree; this reflects some structural change in the translation process. At this point each subtree pair is a transfer rule, but with all items fully specified lexically.

Aligning and processing the entire corpus in this way will yield a large number of such rules. We can then merge these rules automatically. Two rules can be combined if, for each syntactic function in the source language which the rules have in common, the rules specify the same semantic class constraint (or two words which can be generalized to the same semantic class) and the same target language syntactic function. On the other hand, if two such rules specify the same semantic class constraints but different target language syntactic functions, these rules are in conflict; it may be possible to resolve this conflict by creating finer semantic classes and repeating the procedure.

Some languages (notably Japanese) permit arguments of the verb to be freely omitted, in which case (when we are translating into a language which does not permit comparably free omissions) these arguments must be recovered from context. Such recovery, of course, does not fit into the simple transfer rule organization outlined above, which translates sentences separately. We have found, however, that many cases can be handled by constructing sublanguage patterns for verbs (in effect, semantic case frames) in the source language, marking some arguments in these patterns as semantically essential, filling such arguments when they are not present in the source text (with either a generic filler or the most recent entity of that semantic class), and then performing the transfer. In terms of a discovery procedure, however, this raises the question of how such essential arguments may be identified. It may be possible to do so from an analysis of transfer patterns (in which certain arguments are sometimes absent in the source tree but always present in the target tree) or from an analysis of the different semantic patterns of a verb within a single language.

3. Examples

To evaluate the feasibility of this approach, we have been studying a small bilingual corpus, the Primer for the FOCUS Query Language, which is available in both English and a faithful, high-quality Japanese translation. We had previously reported on a sublanguage analysis we had performed of these manuals (Kosaka et al. 1988, Teller et al. 1988), which showed a high level of correspondence between the sublanguage patterns in the two languages which we felt could be the basis for translation (transfer) rules. Since that time we have constructed a small translation system based on these sublanguage patterns which has successfully translated a portion of the Japanese FOCUS Primer into English (Peng 1992); this has given us a more accurate picture of the transfer rules which will be required. We are now studying the sentences from the manual, as analyzed by our Japanese and English analyzers, to see how transfer rules could be automatically acquired.

We have implemented the tree matching procedure described above, and have tested it on a small number of examples, including those shown here. We present here three pairs of sentences, along with tree diagrams which correspond to the regularized syntactic structures produced by our analyzers. These tree diagrams have been simplified to show only the heads of the NP and S structures; the NP and S nodes themselves, along with various syntactic markers such as "tense" and "style", have been omitted.

Figures 1 and 2 show a relatively simple case, where there is a near-perfect match between the structures in Japanese and English. Some nodes of the trees are numbered to show the correspondences between nodes. In most cases, a standard bilingual lexicon should allow us to establish the correspondences between nodes (e.g., "doosi" <==> "verb", "tan=itu" <==> "single"). The correspondence between "motiiru" (literally, "use") and "include" may not be obtainable from the dictionary, but can be established based on the fact that all the operands (daughter nodes in the tree) correspond in the two trees. We can then extract transfer patterns such as "motiiru" + subject "user" + object "doosi" + de "tablecommandnai" --> "include" + subject "user" + object "verb" + in "table command", and then, by comparing it to other examples, generalize this using semantic word classes ***tablecommand-class*** for "verb" and ***location-class*** for "table command".

In the Japanese trees we have assumed that the implicit arguments were identified as semantically essential arguments in Japanese case frames and have therefore been recovered. These recovered arguments are shown enclosed in brackets. In the cases shown here, the recovered argument represents a default subject ("you", "the user", etc.) for particular verbs. We could alternatively have left these arguments unrecovered in the source language tree, and discovered the appropriate fills as part of the transfer rule discovery process. This, however, would not be adequate for cases where the arguments had to be recovered from prior discourse.

Figures 3 and 4 show a more complex example. Clearly the match between nodes is not 1 -to-1 here. We would establish a correspondence between the subtree representing "PRINT, LIST no izureka de siteisita field" and "data with either PRINT or LIST".[2] We could then generalize "PRINT or LIST" to the semantic class ***tablecommand-class***. Also, there is no structure in Japanese corresponding to the English "You will be able to ..." so no transfer rule would be created. (We noted in our earlier study (Teller et al. 1988) that there are sometimes significant differences in the top-level structures between the two languages.)

A third example, shown in Figures 5 and 6, also shows cases where there is a structural mismatch between the two trees, which will lead to structural transfer rules. In particular, we see one of the frequent cases where a prepositional phrase in English is rendered as a relative clause in Japanese: "table command" + with "verb" <==> "tablecommand" + rel ("ni yoru" + object "doosi"). We also see another case where a node correspondence between "syuutokusita" ("learn") and "develop" would not be licensed by the dictionary but is established on structural

[2] The attachment of the "with" modifier is ambiguous in the English sentence, so the procedure would select the parse that is in closest structural alignment with the Japanese parse. That is the analysis we have shown here.

grounds.

4. Prospects

We have suggested here how 'rationalist' and 'empiricist' methods might be combined to produce a more effective development methodology for machine translation systems. As we obtain a deeper understanding of aspects of the language besides syntax, the boundary between what is best hand-crafted and what is best discovered automatically from text may change, but the basic principle for melding these two approaches will remain the same.

Evaluating the approach we have sketched above will involve a rather ambitious program involving several stages of automated acquisition. Our current intent is to approach this incrementally, by developing the semantic classes manually and initially limiting the acquisition process to the semantic co-occurrence constraints (with which we already have considerable experience) and the transfer patterns.

Acknowledgements

This report is based upon work supported by the National Science Foundation under Grants IRI-89-02304 and IRI-89-02269 and by the Defense Advanced Research Projects Agency under Grant N00014-90-J-1851 from the Office of Naval Research.

References

(Baker 1979)

J. K. Baker. Trainable Grammars for Speech Recognition. ***Speech Communication Papers for the 97th Meeting of the Acoustic Society of America,*** D. H. Klatt and J. J. Wolf, editors.

(Brown et al. 1990)

P. F. Brown, J. Coke, S. A. Della Pietra, V. J. Della Pietra, F. Jelinek, J. D. Lafferty, R. L. Mercer, and P. S. Roossin. A Statistical Approach to Machine Translation. ***Computational Linguistics,*** 16 (2).

(Grishman et al. 1986)

R. Grishman, L. Hirschman, and N. T. Nhan. Discovery Procedures for Sublanguage Selectional Patterns: Initial Experiments. ***Computational Linguistics,*** 12 (3): 205-216.

(Grishman and Sterling 1992)

R. Grishman and J. Sterling. Acquisition of Selectional Patterns. To appear in ***Proc. 14th International Conf. on Computational Linguistics (COLING 92),*** Nantes, France.

(Hirschman et al. 1975)

L. Hirschman, R. Grishman, and N. Sager. Grammatically-based Automatic Word Class Formation. ***Information Processing and Management,*** 11 (1/2): 39-57.

(Hirschman 1986)

L. Hirschman, L. Discovering sublanguage structures. In R. Grishman and R. Kittredge (Eds.). ***Analyzing language in restricted domains: Sublanguage description and processing.*** Hillsdale, NJ: Erlbaum, 211-234.

(Kosaka et al. 1988)

M. Kosaka, V. Teller, and R. Grishman. A Sublanguage Approach to Japanese-English Machine Translation. In ***New Directions in Machine Translation,*** D. Maxwell, K. Schubert & A. Witkam, eds., 109-121. Dordrecht: Foris.

(Nagao1987)

M. Nagao. Role of Structural Transformation in a Machine Translation System, in ***Machine Translation,*** S. Nirenburg, editor, Cambridge: Cambridge University Press.

(Peng 1992)

P. Peng. ***Japanese/English Machine Translation Using Sublanguage Patterns and Reversible Grammars.*** Doctoral Dissertation, Computer Science Dept., New York University.

(Sato and Nagao 1990)

S. Sato and M. Nagao. Toward Memory-based Translation. ***Proc. 13th International Conf. on Computational Linguistics,*** Helsinki, 247-252.

(Sumita and lida 1991)

E. Sumita and H. lida. Experiments and Prospects of Example-Based Machine Translation. ***Proc. 29th Annual Meeting Assn. for Computational Linguistics,*** Berkeley, CA.

(Teller et al. 1988)

V. Teller, M. Kosaka, and R. Grishman. A Comparative Study of Japanese and English Sublanguage Patterns. ***Proc. Second Int'l Conf. on Theoretical and Methodological Issues in Machine Translation of Natural Languages.***

(Tsujii 1986)

J. Tsujii. Future Directions of Machine Translation. ***Proc. 11th International Conf. on Computational Linguistics (COLING 86),*** Bonn.

Figure 1. Regularized tree for "You can include more than one verb in a single TABLE command.". Numbers in parentheses indicate correspondence between the nodes in this tree and the tree for the corresponding Japanese sentence, shown in Figure 2.

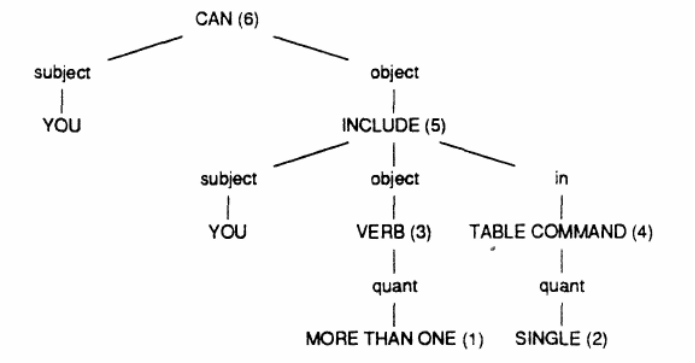

Figure 2. Regularized tree for the corresponding Japanese sentence, "Tan=itu no tablecommandnai de fukusuu no doosi o motiiru koto ga dekimasu.".

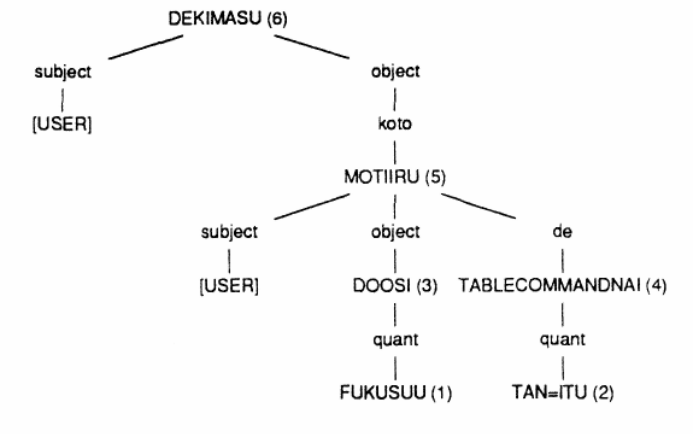

Figure 3. Regularized tree for "You will be able to use an ACROSS phrase to sort data with either PRINT or LIST."

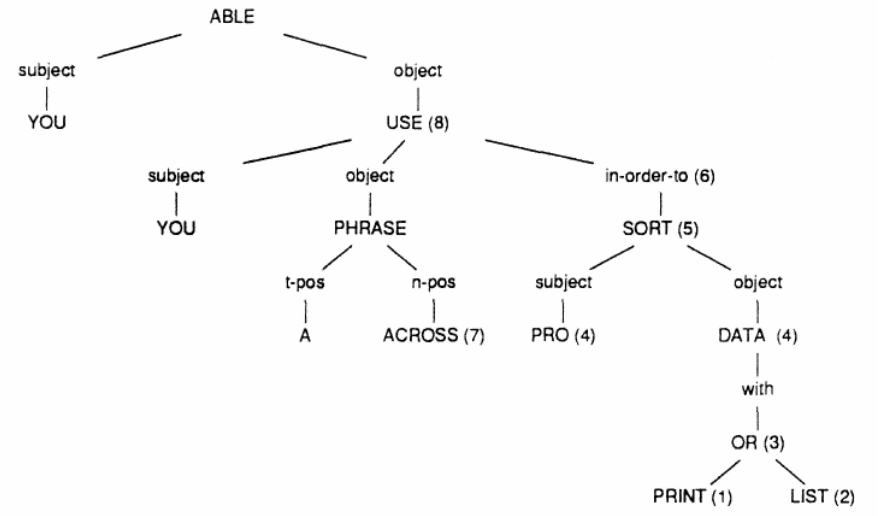

Figure 4. Regularized tree for the corresponding Japanese sentence, "PRINT, LIST no izureka de siteisita field o bun=ruisuru tame no ACROSS no siyoo."

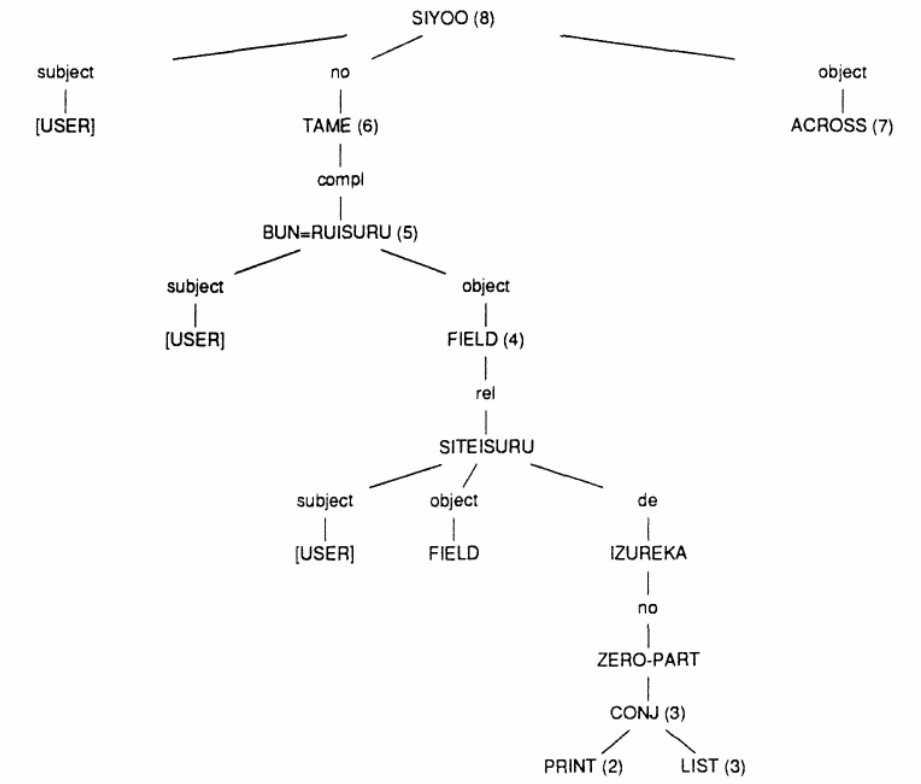

Figure 5. Regularized tree for "Techniques for doing this are discussed in section 8 on page 173 after a complete development of the TABLE command with a single verb."

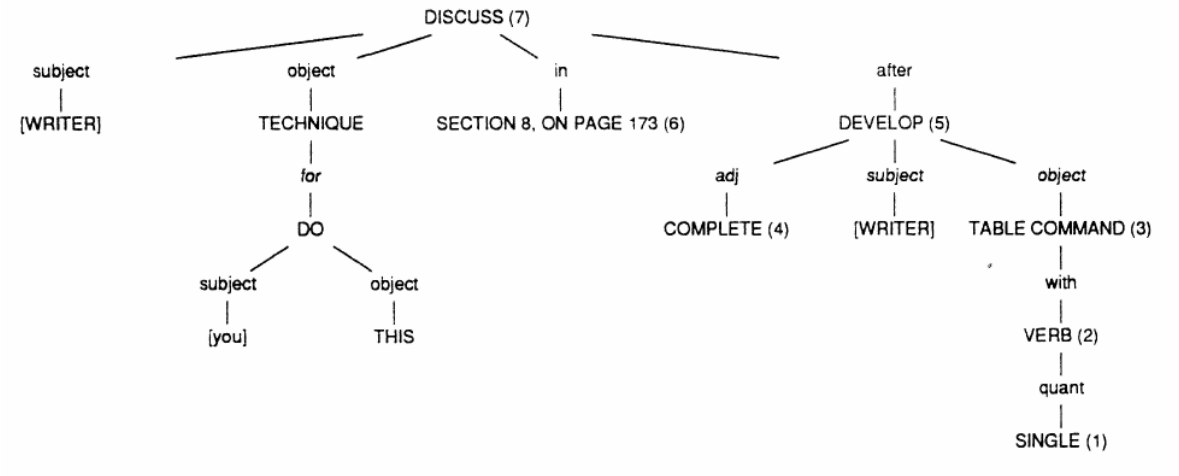

Figure 6. Regularized tree for "Kono syori no gijututeki na naiyoo wa tan=itu doosi ni yoru tablecommand o kanzen ni syuutokusi ta ato de dai 8 shoo 167 page de setumeisi masu."