A COMPARATIVE STUDY OF JAPANESE AND ENGLISH SUBLANGUAGE PATTERNS

Virginia Teller
Hunter College and the Graduate Center
The City University of New York

Michiko Kosaka
Monmouth College

Ralph Grishman
New York University

ABSTRACT

As part of a project to develop a Japanese-English machine translation system for technical texts within a limited domain, we conducted a study to investigate the roles that sublanguage techniques (Harris, 1968) and operator-argument grammar (Harris, 1982) would play in the analysis and transfer stages of the system. The data consisted of fifty sentences from the Japanese and English versions of the *FOCUS Query Language Primer,* which were decomposed into elementary sentence patterns. A total of 187 pattern instances were found for Japanese and 191 for English. When the elements of these elementary sentences were classified and compared with their counterparts in the other language, we identified 43 word classes in Japanese and 43 corresponding English word classes. These word classes formed 32 sublanguage patterns in each language, 29 of which corresponded to patterns in the other language. This paper examines in detail these correspondences as well as the mismatches between sublanguage patterns in Japanese and English.

The high level of agreement found between sublanguage categories and patterns in Japanese and English suggests that these categories and patterns can facilitate analysis and transfer. Moreover, the use of operator-argument grammar, which incorporates operator trees as an intermediate representation, substantially reduces the amount of structural transfer needed in the system. A pilot implementation is underway.

This research was supported in part by the Defense Advanced Research Projects Agency under Contract N00014-85-K-0163 from the Office of Naval Research.

1. Introduction

For a pair of disparate languages -- Japanese and English -- we are developing a machine translation system based on a sublanguage analysis of technical texts within a restricted domain. As developed by Harris (1968), the sublanguage approach to linguistic analysis entails delimiting a circumscribed domain of discourse, selecting sample texts in the domain, and identifying the word classes and patterns of word class co-occurrences that are specific to the sublanguage. Sublanguage patterns provide important benefits in both the analysis and transfer stages of a machine translation system. During analysis they serve to block incorrect parses and aid in the recovery of elided material. This recovery is particularly important in a language like Japanese, where zeroing is far more widespread than in English. The use of sublanguage patterns in the transfer phase rests on the premises that (1) there is a correspondence between the sublanguage categories and patterns in the source language (Japanese) and the target language (English); and (2) these categories and patterns are the appropriate units for lexical disambiguation. In addition, the operator-argument grammar framework (Harris, 1982) that we have adopted, which incorporates operator trees as an intermediate representation, further explicates the underlying relationships among sublanguage word classes and substantially reduces the amount of structural transfer needed in the system.

The sublanguage approach has found several computational applications, in North America, particularly in the work of the Linguistic String Project at New York University (e.g. Sager, 1981) and the TAUM group at the University of Montreal, where sublanguage grammars have been used in machine translation projects (Lehrberger, 1982; Isabelle & Bourbeau, 1985; Kittredge, 1987). To date, however, these techniques have not been tested on languages as dissimilar as Japanese and English, and the correctness of the premises outlined above is far from assured. The close correspondence between French and English sublanguage patterns found by the TAUM group is not guaranteed to carry over to Japanese and English. The relationships could just as easily be one-to-many or many-to-one. We have investigated this question with the goal of using sublanguage categories and patterns to facilitate the computer analysis of source texts in Japanese in the sublanguage domain of computer manuals intended as instructional material. Our efforts have concentrated on the *FOCUS Query Language Primer,* which has been published in both Japanese and English.

On the basis of a comparative linguistic analysis of Japanese and English using Harris's operator-argument framework, we have proposed a novel design for a machine translation system (Kosaka, Teller & Grishman, 1988). A central claim of our proposal is that this model, which is essentially a transfer system without a component for structural transfer, offers a middle road between the transfer and interlingua approaches to machine translation. Since the strength and validity of this claim rest squarely on the results of our linguistic analysis, most of this paper is devoted to a detailed description of that analysis and an assessment of the significance and implications of the results for machine translation.

2. Comparative linguistic analysis

2.1 Method. The distributional analysis of sublanguage texts according to the principles of operator-argument grammar produces a set of sublanguage word classes and a set of word class co-occurrence patterns. The co-occurrence constraints embodied in these patterns are viewed as a manifestation of the underlying semantic constraints of the domain. The patterns that emerged from our study were obtained in a two step process. First, each sentence in the sample texts was decomposed into its constituent elementary sentences. This process regularizes surface representations in order to arrive at canonical representations that accord with information content. For example, the sentence *IN-GROUPS-OF and TOP can be used with ACROSS,* contains four elementary sentences.:[1]

(1) a. *U* uses IN-GROUPS-OF with ACROSS.
b. *U* uses TOP with ACROSS.
c. S1 and S2.
d. can S.

In the second step these elementary sentences were classified into operator-argument co-occurrence patterns. The operators that occur in a sublanguage fall into four classes. Zero-order operators, which include most nouns, accept no arguments. First-order operators, which take zero-order operators as arguments, comprise the operators that appear in base sublanguage relationships (kernel sentences). These operators include the verbs in subject-verb-object patterns, and their arguments are the subject and object word classes permitted by the sublanguage. The class of second-order operators contains certain modifiers such as modals as well as disjunction, coordinate and subordinate conjunctions, etc. whose arguments are first-order operators (i.e. kernel sentences). Operators that produce paraphrases (e.g. passive, nominalization) also belong to this class. The fourth class, meta-operators, consists in our corpus of verbs that belong to the sublanguage of instructional material rather than the domain of computer manuals. These include "mental" verbs such as *hope, learn, observe* and *understand* as well as *discuss, explain, introduce* and *present,* which take human subjects.

Co-occurrence patterns in operator-argument format are labeled with the word class of the operator followed by the word class of the arguments. Sentences (la) and (1b), for example, are instances of the kernel pattern USE_WITH-USER-OPTION-PHRASE, which consists of a first-order operator with zero-order arguments. (Note that the pattern specifies USER as the subject argument that is missing but understood in the elementary sentence.) Sentence (1c) falls into the class AND/OR-Sl-S2-(Sn), and (1d) is a member of MODAL-S. Both of these patterns contain second-order operators with kernel sentences as arguments. Six word classes are also illustrated: OPTION (with members IN-GROUPS-OF and TOP), PHRASE (with member ACROSS), USER, USE-WITH, AND/OR, and MODAL.

Fifty sentences were selected for analysis from a twenty page section of the Japanese and English versions of the FOCUS manual. These source and target texts gave us an independent standard by which to judge our techniques and results, thereby eliminating the need for translation on our part and the possibility of bias that could be introduced if we translated a particular text ourselves. Working independently, two linguists listed the co-occurrence

[1] IN-GROUPS-OF, TOP, and ACROSS are keywords in the FOCUS query language. The symbol *U* stands for "unspecified", that is, a missing or zeroed argument.

patterns for each sentence. These included elementary sentences with higher order operators such as coordinate and subordinate conjunctions as well as subject-verb-object structures and prepositional/postpositional phrases that exhibited selectional restrictions specific to the domain.

2.2 Results. A total of 187 pattern instances were found for Japanese and 191 for English. When the elements of these elementary sentences were classified and compared with their counterparts in the other language, we identified 43 word classes in Japanese and 43 corresponding English word classes. These word classes formed 32 patterns in Japanese and 32 in English that occurred more than once. Twenty-nine of the Japanese patterns correspond to English patterns in the sense that they have identical argument structures and convey the same meaning. This was an encouraging outcome given the possible number of combinations of 43 word classes that could appear in kernel patterns consisting of two, three, and even four elements.

Table 1 lists the 43 word classes that emerged from our analysis. Approximately half of the classes contain only one member, owing to the relatively small size of the sample texts. The largest class, TABLE, which consists of all the keywords and fields in the FOCUS TABLE command, comprises over a dozen members. Examples of robust classes include the verb class COMPUTE with members *{gookei-suru, group-wake-suru, keisan-suru, sansyutu-suru, syuturyoku-suru,...}* in Japanese and *{count, compute, generate, group, sum, ... }* in English, and the noun class VALUE, whose Japanese and English members are *{atai, gookei, kekka, suuti]* and *{result, summary, total, value},* respectively. Words that occur in different contexts are considered homographs and are assigned to more than one word class. Examples are *syuturyoku-suru* and *generate,* which belong to two verb classes: CREATE (as in *generate a report)* and COMPUTE (as in *generate subtotals).* There are also two classes labeled IN and three with the label USE. These are operators that appear in two or more patterns with different argument structures.

Table 1. Japanese-English word classes.

zero-order	first-order		second-order	meta
DBASE	COMBINE	REQUIRE	AFTER	MEAN
FIELD	COMPONENT	SPECIFY	AND/OR	META
FOCUS	COMPUTE	USE1	BEFORE	SAME
FORMAT	CREATE	USE2	IF	
HUMAN	DISPLAY	USE3	IN2	total: 3
OPTION	FIT	USEFUL	IN-ORDER-TO	
PHRASE	GROUP	USE-WITH	MODAL	
REPORT	IN1	WRITE	NEG	
TABLE	PRINT		PERFORM	
USER			RELATE	
VALUE	total: 17		WHEN	
VERB				
			total: 11	
total: 12				

Table 2. Japanese-English sublanguage patterns. The frequencies in Japanese and English and the number of matching occurrences for each pattern are given in brackets. Superscripts refer to numbered commentary in the text.

kernel sentence patterns	higher order operator patterns[4]
COMBINE-USER-TABLE [4,3,3]	AFTER-S1-S2 [1,1,1]
COMPONENT-TABLE [3,3,3]	AND/OR-Sl-S2-(Sn) [14,13,10]5
COMPUTE-FOCUS-VALUE [9,9,8]	BEFORE-S1-S2 [4,4,4]
CREATE-TABLE-REPORT [5,4,4]	IF-S1-S2 [2,1,1]
DISPLAY-VERB-VALUE [3,0,0][1]	IN2-USE2-FIELD [4,4,4]
DISPLAY-FOCUS-VALUE [4,1,1][1]	IN_ORDER_TO-S1-S2 [3,4,3]
FIT-VALUE-FORMAT [2,2,2]	MODAL-S [8,7,5][6]
GROUP-FOCUS-FIELD [2,1,1]	NEG-S [4,3,3]
IN1-PHRASE-TABLE [2,2,2]	PERFORM-FOCUS-COMPUTE [8,6,6][7]
PRINT-FOCUS-NP [3,3,2]	RELATE-VERB-COMPUTE [2,1,1]
REQUIRE-FOCUS-NP [4,4,4]	WHEN-S1-S2 [5,4,3]
SPEC-USER-PHRASE [2,0,0][2]	
USE1-USER-DBASE [2,0,0][3]	total: 11
USE2-USER-TABLE [26,17,16][3]	
USE3-USER-VALUE [2,1,1]	
USEFUL-VERB-REPORT [3,2,2]	meta-operator patterns
USE_WITH-USER-OPTION-PHRASE [2,2,2]	
WRITE-USER-TABLE [2,2,2]	MEAN-NP-NP [5,5,5]
	META-HUMAN-X [17,14,7][8]
total: 18	SAME-NP-NP [2,2,2]
	total: 3

Table 2 shows the 32 elementary sentence patterns that occurred more than once in Japanese.[2] The numbers in brackets give the frequency in Japanese, the frequency in English, and the number of matching occurrences for each pattern. A matching occurrence is one where corresponding Japanese and English sentences contain Instances of the same sublanguage pattern.

The kernel sentence patterns define a set of base relationships among word classes that constitute a partial description of the domain knowledge. These patterns, together with the higher order and meta-operator patterns, embody a set of semantic constraints that can be stated as selectional restrictions on word class co-occurrences, for example, on the subject and object word classes allowed with a particular class of verbs. During the analysis phase of machine translation sublanguage patterns serve to reduce the ambiguity of the source language text and block incorrect parses proposed on the basis of syntactic information alone. As discussed in Kosaka, Teller and Grishman (1988), these patterns also make it possible to resolve ellipsis. Our intention in performing a linguistic analysis of both English and Japanese, however, was to determine the degree to which sublanguage patterns could

[2] The pattern AFTER-S1-S2 is included because of the importance of ato-de/after as a subordinate conjunction, even though there was only a single instance in the corpus.

Table 3. Instances of the CREATE-TABLE-REPORT kernel sentence pattern. Page references are given for Japanese and then English, followed by the sentence number. An asterisk indicates a matching pattern occurred in the other language.

JAPANESE:

*p50/p56 s2: O-ga report-o sakusei-suru.
*p55/p61 s4: TABLE-command-ga report-o sakusei-suru.
*p56/p62 s4: TABLE-command-ga report-o sakusei-sita.
*p62/p68 s1: TABLE-command-ga report-o sakusei-suru.
p66/p72 s4: O-ga report-o syuturyoku-suru.

ENGLISH:

*p50/p56 s1: U create report.
*p55/p61 s4: TABLE-commands produce reports.
*p56/p62 s4: TABLE-commands generate reports.
*p62/p68 s1: TABLE-commands produce reports.

play a role during the transfer phase as well, by providing a principled basis for translating source language vocabulary and semantic content into equivalent terms in the target language. For this purpose the numbers in brackets in Table 2 are of crucial importance. These numbers indicate the strength of the match between the patterns found in Japanese and those in English and hence the degree of similarity that can be expected between operator-argument trees in the two languages.[3]

Table 4. Instances of the IN2-USE2-FIELD kernel sentence pattern. Page references are given first for the Japanese, then the English text. An asterisk indicates a matching pattern occurred in the other language.

JAPANESE:

*p51/p57 s5: mokuteki-field-tyuu de no siyoo.
*p51/p57 s7: mokuteki-field-tyuu de no siyoo.
*p51/p57 s7: mokuteki-field-tyuu de no siyoo.

ENGLISH:

*p51/p57 s5: use in object list.
*p51/p57 s7: use in object list.
*p51/p57 s7: use in object list.

[3] Not shown are the English elementary sentence patterns that had no equivalents in Japanese; only three of these occurred more than once. Certain types of prepositional phrases are also omitted from the analysis.

Within the domain of texts we have examined so far, the correspondences between Japanese and English sublanguage patterns are excellent. The highest agreement lies in the group of 18 kernel sentence patterns. Mismatches are more common in the patterns with higher order and meta-operators. The largest proportion of discrepancies is accounted for by just two patterns, USE2-USER-TABLE and META-HUMAN-X (where X stands for NP or S). Tables 3 and 4 illustrate nearly perfect matches between Japanese and English for two major syntactic relations. The CREATE-TABLE-REPORT kernel sentence pattern describes a subject-verb-object structure, while IN2-USE2-FIELD incorporates a postpositional/prepositional phrase. In the following section an explanation is given for each entry in Table 2 where there is a discrepancy of more than one between the number of matching occurrences (the third item in brackets) and the frequencies in Japanese and English (the first and second numbers in brackets). The commentary is keyed to superscript references in the table.

3. Commentary on the operator-argument patterns

[1] The absence of the pattern DISPLAY-VERB-VALUE in English is explained by a construction with *allow* for which there was no equivalent in Japanese. An example is the sentence *SUM and COUNT allow you to display results,* which decomposes into four elementary sentence patterns: SUM allows S; COUNT allows S; you display results; S1 and S2. The first argument in the English version of the pattern with DISPLAY is USER (i.e. *you)* instead of VERB (i.e. *SUM, COUNT).* In the case of the three DISPLAY-FOCUS-VALUE mismatches, the English text used the verbs *appear, print,* and *report,* which are not members of the DISPLAY class.

[2] Although there was no equivalent for the SPEC-USER-PHRASE pattern in English, an acceptable English translation can be generated from the Japanese pattern after lexical transfer.

[3] Instances where a subordinate clause in Japanese was expressed as a prepositional phrase in English account for the discrepancies between Japanese and English in the patterns with the USE class of verbs. In the next section an example of the USE1-USER-DBASE pattern is examined in which the preposition *from* is used in English. The preposition most commonly found instead of the USE2-USER-TABLE pattern is *with.*

[41 Omitted from the analysis are five quantifier word classes and five sublanguage patterns with quantifier operators. Since quantifiers can modify almost any NP, they impose very little selectional specificity on their arguments and therefore play a limited role in providing sublanguage co-occurrence restrictions. Although the match between Japanese and English quantifier patterns was excellent ([13,12,11] in terms of bracketed numbers), the analysis of quantifiers in Japanese is complicated and lies beyond the scope of this paper.

[5] Differences in the distribution of conjunctions arose in situations where either the Japanese or the English text conjoined two sentences that appeared as two separate sentences in the other language.

[6] The mismatches observed in the pattern MODAL-S were due primarily to cases where Japanese used modals for politeness, the result being sentences that convey quite different meanings in Japanese and English.

[7] The pattern PERFORM-FOCUS-COMPUTE appeared in one Japanese sentence where the English version used the verb *involve* Instead of *perform* and in one sentence with a completely different structure in English.

[81 The symbol X stands for 'NP or S'. The meta-operators in our sample revealed a particularly Interesting source of deviation between Japanese and English. In many cases sentences containing members of this class of operators express very different meanings in the two languages. Example (2b) below gives a literal rendering in English of sentence (2a) from the FOCUS manual, while (3) gives the actual sentence that appeared in the English version:

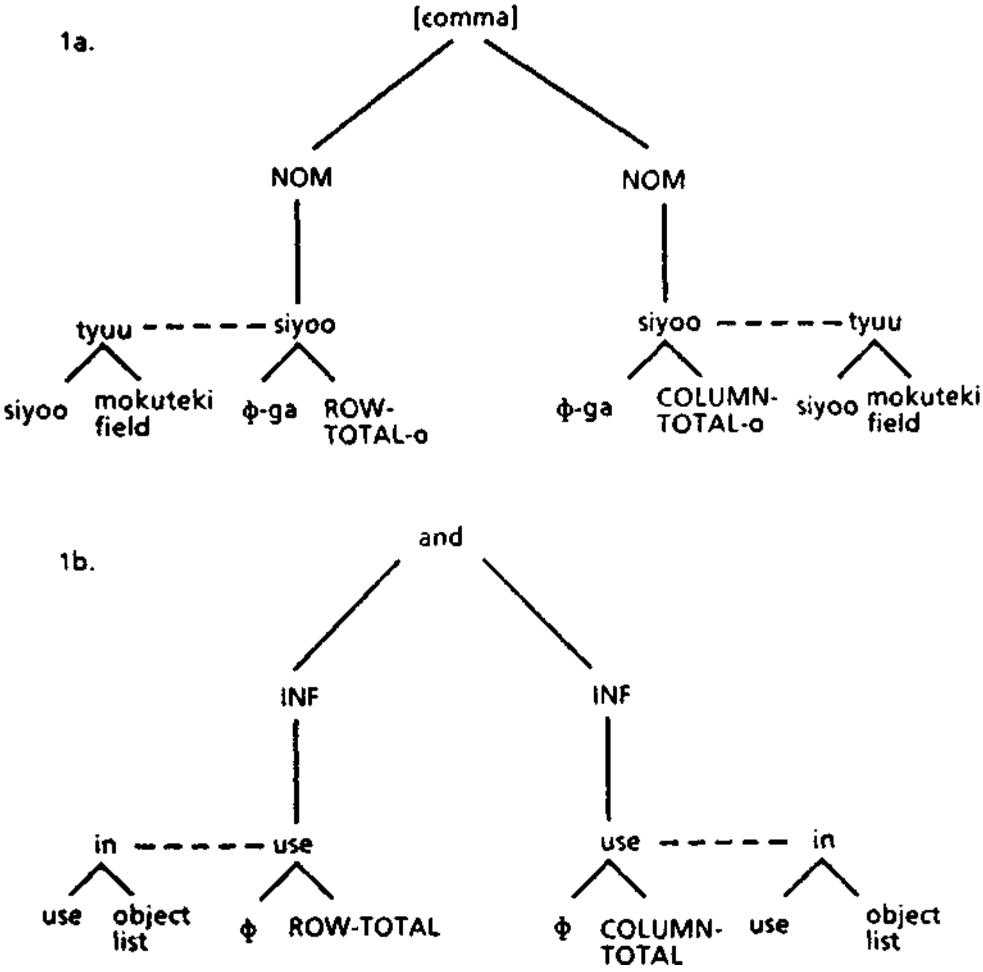

Figure 1. Operator trees for examples (4a) and (4b). Dashed lines indicate an adjunct relationship to the parent node. Solid lines indicate an operator-argument relationship.

(2) a. Kokomade de kaki-no TABLE command no kakukoomoku-ga rikai-dekita-to omoimasu.
b. By now we hope that you are able to understand each component of the following TABLE command.

(3) At this point the following components of the TABLE command have been introduced.

The meta-operators appear to be one of the parameters that contribute to stylistic differences in expression between the two languages.

4. Implications for machine translation

Within the subdomain of texts we have examined so far, the correspondences between Japanese and English are not limited to sublanguage word classes and co-occurrence patterns but extend to the overall structure of operator trees as well. For example, the Japanese nominalization (4a) and its matching English infinitive clause (4b) are represented in our operator tree system as shown in Figures la and 1b, respectively:

(4) a. Mokuteki field tyuu de no ROW-TOTAL, COLUMN-TOTAL no siyoo.
b. Use ROW-TOTAL and COLUMN-TOTAL in the object list.

Structurally the trees are identical, but the nominalization operator appears in Japanese where English uses an infinitive. Identical operator-argument patterns also appear in the two trees — the kernel sentence patterns USE2-USER-TABLE and IN2-USE2-FIELD and the paraphrastic operator pattern AND/OR-S1-S2-(Sn). Although the USE2 operator *siyoo/use* allows two arguments, the O's indicate that no argument appears in subject position in either Japanese or English.

The strong similarities between Japanese and English operator trees suggest that, with operator trees as an intermediate representation, it may be possible to construct a system to translate Japanese into English without the structural transfer usually associated with such systems. A successful translation of (4a) into (4b) can be achieved solely on the basis of lexical transfer. No restructuring is necessary at the operator-argument level of analysis.

As for the data not accounted for in this manner, several options are available to the designer of an MT system. When no match is found for a source language pattern, the system could fail to produce a translation, restructure the tree into a comparable target language pattern, or proceed with lexical transfer without restructuring. We have adopted the last of these options. Our strategy has been to assess whether an acceptable English sentence could, in principle, be generated from the Japanese operator tree. Although devices could be introduced to map the Japanese sublanguage patterns with no equivalent in English into different, and possibly more appropriate, operator-argument structures, we prefer to maintain the position of avoiding such structural change as long as the Japanese operator tree can be used as the basis for a grammatical English sentence. This tactic has proven successful in the majority of discrepancies we have encountered so far.

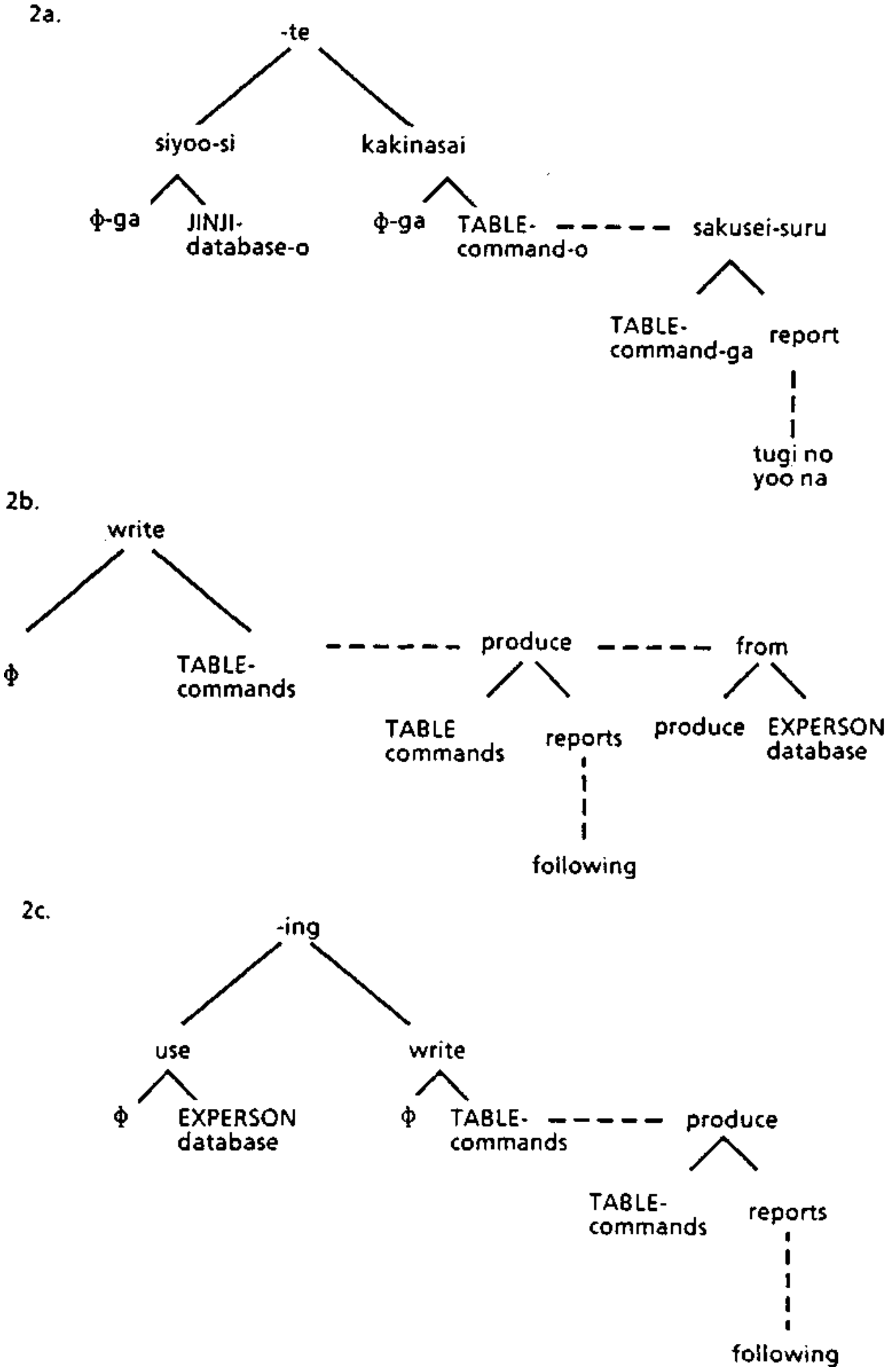

Figure 2. Operator trees for sentences (5a), (5b) and (6). Dashed lines indicate an adjunct relationship to the parent node. Solid lines indicate an operator-argument relationship.

Sentences (5a) and (5b) below illustrate how the operator-argument level of intermediate representation allows a graceful recovery from an apparent mismatch in sublanguage patterns:

(5) a. JINJI database o siyoo site tugi no yoo na report o
sakusei suru TABLE command o kakinasai.
b. Write TABLE commands that will produce the following
reports from the EXPERSON data base.

Although, as shown in Figures 2a and 2b, the Japanese and English versions share instances of the kernel sentence patterns WRITE-USER-TABLE and CREATE-TABLE-REPORT, the Japanese sentence also contains the pattern USE1-USER-DBASE, which is lacking in English. In addition, the subordinate conjunction *-te* introduces a clause in Japanese which is expressed in English as a prepositional phrase with *from.* Rather than restructuring the Japanese operator tree into one that resembles the English version, we can obtain an acceptable translation using only lexical transfer. The result, shown in Figure 2c, will produce a sentence like (6):

(6) Using the EXPERSON database, write TABLE commands that will
produce the following reports.

5. Implementation

The results of our indicate that the sublanguage approach is worth pursuing for the analysis and lexical transfer stages of a machine translation system. In parallel with our sublanguage studies we have begun implementation of a pilot MT system. We have taken a previously developed question-answering system and incorporated a small, core Japanese grammar and regularization component capable of parsing and producing operator trees for simple sentences. The parser has been coupled to a semantic analyzer that utilizes sublanguage patterns and an existing retrieval component to produce a Japanese version of the question answerer. We then added a lexical transfer component based on the same sublanguage patterns and an English sentence generator to complete the pilot translation system. Relativization and quantification are among the features of the current implementation, as shown by the following examples of input (7) and output (8):

(7) a. Jane ga A o totta kamoku wa nan desoo ka?
b. Subete no gakusei wa V11 o totta ka?

(8) a. What is the course that Jane got an A in?
b. Did all students take V11?

These examples illustrate two functions that the sublanguage patterns perform in the translation process. First, the Japanese verb totta participates in two sublanguage patterns in this domain, and the two corresponding English patterns involve different verbs *(take* and *receive);* in this way the patterns guide lexical transfer. Second, since Japanese provides no overt marker of which argument is omitted in a relative clause, the sublanguage pattern is required in order to identify the omitted argument in Japanese and pair it with the corresponding argument in the English pattern. In examples such as these, where the missing argument in English is marked by a preposition, the preposition must be generated as part of the English relative clause.

REFERENCES

Harris, Z. 1968. *Mathematical structures of language.* New York: Wiley Interscience.

Harris, Z. 1982. *A grammar of English on mathematical principles.* New York: Wiley.

Isabelle, P. and Bourbeau, L. 1985. TAUM-AVIATION: Its technical features and some experimental results. *Computational Linguistics* 11:18-27.

Kittredge, R. 1987. The significance of sublanguage for automatic translation. In Nirenburg (Ed.), pp. 59-67.

Kittredge, R. and Lehrberger, J., Eds. 1982. *Sublanguage: Studies of language in restricted semantic domains.* Berlin-New York: Walter de Gruyter.

Kosaka, M., Teller, V. and Grishman, R. 1988. A sublanguage approach to Japanese-English machine translation. To appear in *Proceedings of the International Conference on New Directions in Machine Translation.* Dordrecht: Foris.

Lehrberger, J. 1982. Automatic translation and the concept of sublanguage. In Kittredge and Lehrberger (Eds.), pp. 81-107.

Nirenburg, S., Ed. 1987. *Machine translation: Theoretical and methodological issues.* Cambridge: Cambridge University Press.

Sager, N. 1981. *Natural language information processing: A computer grammar of English and its applications.* Reading, MA: Addison-Wesley.