IV- The Linguistic Model : General Aspects

John Lehrberger

Research group in automatic translation

TAUM

Université de Montréal

February 1981

TABLE OF CONTENTS

1. PURPOSE OF SYSTEM

In the period 1976-1980 TAUM has developed a system for machine translation of texts from English to French. The system is intended to be fully automatic in the sense that no human intervention is required during the process from the submission of the original text in English to the delivery of the corresponding French text in normal readable form. Pre-edition is incorporated in the system; it is handled by the machine with no human aid.

The system was designed to translate texts of a specific type, namely aircraft maintenance manuals, rather than arbitrary texts from any field (see section 2-D). However, this does not mean that a new system must be designed to accomodate each change in subject matter to be translated.

The specialized use of language that characterizes certain domains gives rise to what may be called "sublanguages", particularly within technical fields. Since the chief differences between many technical sublanguages are in vocabulary and semantic range (i.e., range of meanings of individual words), the system should be able to accomodate changes in vocabulary and semantic features without changing the overall design. This would permit adaptation for use say in various technical fields without having to build an entire new grammar every time a new subject matter is to be translated. In view of these facts, the TAUM system aims for maximum generality while taking advantage of the restrictions present in texts within particular fields.

2. LINGUISTIC ASSUMPTIONS

A. Input texts are grammatically correct

It is assumed that the input to the system consists of texts that are grammatically correct. The system is not designed to correct bad writing, but to translate acceptable texts. This assumption has important consequences since the parser can then make predictions about the structure of a sentence as it is being parsed which would not be possible otherwise. The processing of arbitrary strings over a given vocabulary, separating sentences from non-sentences and determining the structure of the former, would be a considerably more difficult enterprise than determining the structure of strings that are known in advance to be correct sentences.

B. Sentences are unambiguous in context

It is also assumed that the sentences of a text are not ambiguous in context, even though they may be ambiguous when taken in isolation. For example, one assumes that an aircraft maintenance manual does not give ambiguous instructions to a mechanic. Thus it might be preferable to have a means of choosing the most likely interpretation whenever it seems that more than one is possible, rather than producing multiple outputs.

Ambiguities may result from homography: a "word" may belong to more than one grammatical category (filter can be a noun or a verb), or it may have different senses within the same grammatical category (as a noun, attachment can signify the action of attaching or a physical object which is attached to something).

Ambiguities may also exist even when the individual words are not ambiguous, as when more than one bracketing of a string of words is possible:

(1) [defective] [priority valve] versus [defective priority] [valve]

(2) [hydraulic system] [No.1 reservoir] versus [hydraulic system No.1] [reservoir].

(3) [more] [widespread damage] versus [more widespread] [damage].

The system is designed to resolve such ambiguities by both semantic and syntactic means. For the moment we simply note that the assumption of non-ambiguity leads to a model which attempts to minimize multiple outputs - and one means of accomplishing this is to put as much "knowledge of the world" as possible into the grammars and dictionaries of the system.

C. Syntax and semantics are not separated

It is assumed that both syntactic and semantic information must be available at all times in order to understand a text. This is reflected in the linguistic model by the fact that syntax and semantics are not separated. Semantic information is coded in the form of semantic features which represent semantically defined subclasses of words. These subclasses are not arbitrary; they are the ones required (in addition to the usual categories such as noun, verb, etc.) to describe various linguistic relations within texts from a given field: which words can "modify" which other words in specific ways, which words can be

grouped together to form phrases of certain types, etc. These subclasses are not the same for texts in all fields. Some are highly specific for a given subject matter (e.g., a class such as igneous might be useful in geology, but not in electronics), while others such as abstract and concrete are likely to be useful in texts from many different fields. Semantic subclasses form a link between subject matter and syntax - between knowledge of the world and grammar.

Each dictionary entry is a composite of the syntactic and semantic properties of the word. Thus the entry for the word pour indicates that:

(4) (a) pour is a verb,

(b) it can be used transitively (He poured the oil into the tank) or intransitively (The oil poured into the tank), and the direct object of the verb in its transitive occurrence (the oil) is the same as the subject of the verb in the corresponding intransitive occurrence,

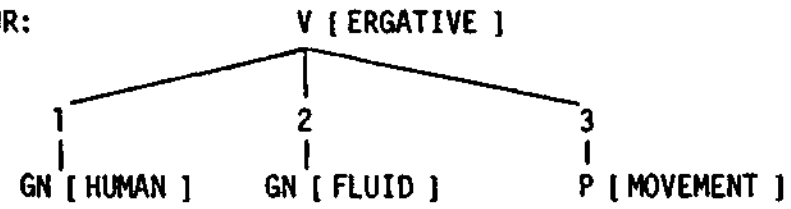
(c) human beings do the pouring,

(d) fluids can be poured, and

(e) the pouring can be into, onto, from or out of something.

The information in (a) and (b) is considered syntactic while that in (c), (d) and (e) is semantic. The actual dictionary entry, which is a string of symbols, may be represented in "tree" form as follows:

(5) POUR:



V, GN and P stand for verb, noun phrase and preposition respectively; [ERGATIVE] signals to the parser that the verb is of the type described in (4-b), thereby eliminating the need for separate dictionary entries for the transitive and intransitive uses (see section 5-d for further discussion); the first two branches of the tree convey the information in (4-c) and (4-d); the third branch encodes (4-e), where P [MOVEMENT] signals the parser that if a prepositional phrase complement occurs with the verb, the preposition itself is of the type used to indicate movement.

At each stage of analysis both syntactic and semantic information are applied and transfer algorithms likewise make use of both syntactic and semantic properties and relations.

D: Texts are restricted in subject matter

From our experience it appears that within certain fields texts are sufficiently restricted in vocabulary, semantic range of individual words and, to some extent, in syntax that automatic parsing of such texts is feasible. This matter is discussed in detail in "Automatic Translation and the Concept of Sublanguage", Lehrberger, J., AILA, 1978.

3. GROSS STRUCTURE

A. Three Stages: analysis, transfer, generation

In broad outline, the system consists of three stages - analysis, transfer and generation - as indicated in Figure 1.

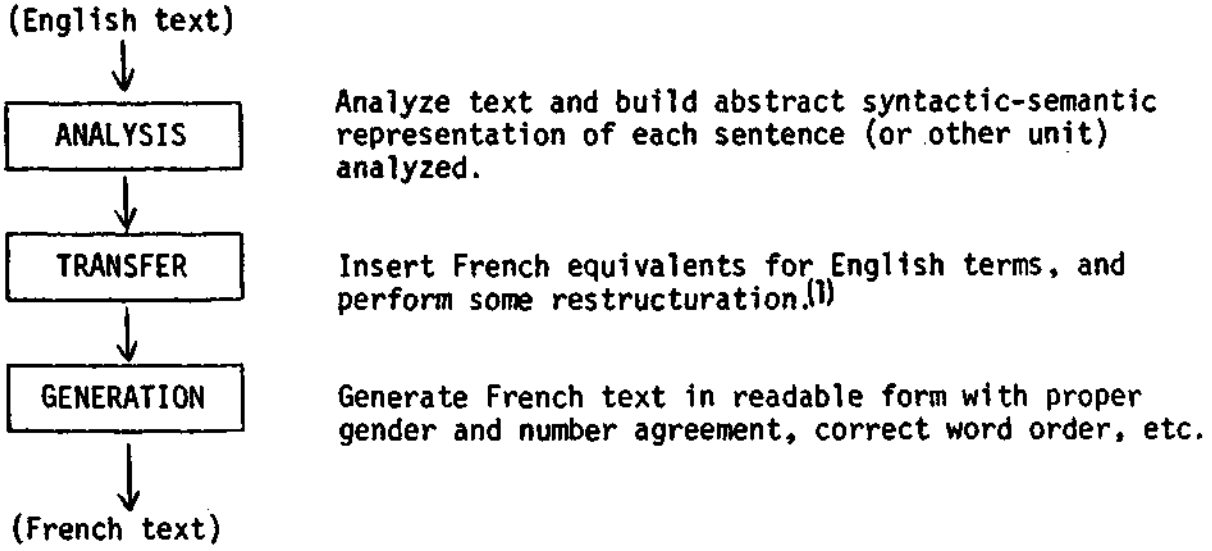


FIGURE 1

B. Justification

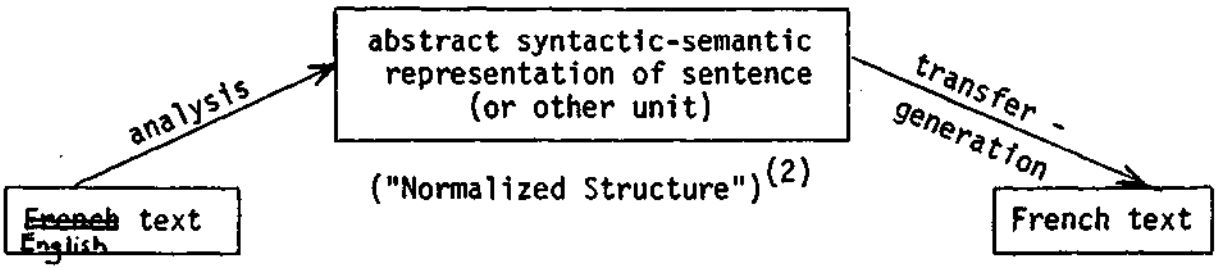
What is the reason for a three stage model of this sort rather than a model, for example, in which there are simply context sensitive rules for replacing English terms with French equivalents without building up an abstract representation of each sentence to serve as input to a transfer stage?

---

(1) To illustrate what is meant by 'restructuration', wooden box (adjective + noun) in English becomes boîte en bois (noun + preposition + noun) in French, which is a change in grammatical structure that accompanies the insertion of French equivalents.

Experience has shown that in order to translate a text it is necessary to understand it. But "understanding" entails access to a complex of semantic and syntactic information, as discussed above. It is the function of analysis to build a structure in some standard form so that transfer can systematically check for the information needed to establish French equivalents. Transfer algorithms can then be based on a uniform representation of the English sentence with syntactic-semantic information in a readily accessible form. This abstract representation of the sentence is of central importance in the model. In fact, the gross structure of the model shown in Figure 1 might, from the point of view of theoretical linguistics, better be represented as in Figure 2.

FIGURE 2



Such a model is well suited for the purpose, stated in Section 1, of obtaining maximum generality. The abstract representation of the sentence is independent of the particular texts to be translated and, furthermore, it is to some extent independent of the source and target languages (see Section 5).

---

(2) There is a systematic ambiguity in the use of the term normalized structure: (1) the particular representation of the structure of a sentence, which constitutes the output of analysis, and (2) the class of structures to which both the output of analysis and the result of each restructuration made by transfer must conform. Thus for each sentence there is a sequence of "tree" structures from the output of analysis through the output of transfer, all conforming to normalized structure in the second sense.

4. UNIT OF TREATMENT: THE SENTENCE

One usually thinks of translation in terms of "finding the right word". However, as every translator knows, a "word" in the source language may have many equivalents in the target language; the correct equivalent is found only after considering the word's role - syntactic and semantic - in the sentence in which it occurs. Therefore the primary focus is on the sentence as the unit to be translated rather than the word.

Although the relations between the words within a sentence are of primary importance in translation, it is also necessary in some cases to make use of information from the surrounding text. This is especially true in the case of tables and charts. If a table contains a column headed PARTS NOMENCLATURE or COMPONENT then the parser must be given a signal that permits it to accept units consisting of noun phrases rather than sentences; if there is a column headed LOCATION, prepositional phrases must be accepted; but in a trouble-shooting chart columns headed PROBABLE CAUSE, ISOLATION PROCEDURE and REMEDY normally contain sentences - or phrases that can be construed as sentences (clogged filter = filter is clogged). In the light of these facts, the system permits the parser to switch to a mode of operation that permits acceptance of units other than sentences while processing a certain segment of text (table, chart, etc.); otherwise, the sentence is the unit to be parsed.

There are also certain intersentential links that affect the translation of particular words. For example, a pronoun in one sentence may refer to a noun in another sentence and require agreement in gender and number with that

noun. Strategies for locating referends of pro-words in other sentences are not very reliable, but fortunately this problem is relatively infrequent in technical manuals. In order to make use of a text based parser rather than one which is sentence based it would be necessary to formalize all intersentential relations, i.e., to write a "text grammar". Linguists have only recently addressed themselves to this task and there are few results that could be incorporated in an automatic parser at present. Furthermore, implementation of a text grammar in the parser (assuming that one could be written) raises some difficult questions concerning methods for making a variety of information from each previously parsed sentence available when parsing other sentences at a distance in the text.

From the point of view of automatic translation the most important relations to understand are those within the sentence, and much work remains to be done in formalizing those relations. The TAUM System therefore takes the sentence as the unit of treatment.

5. BASIC REPRESENTATION OF THE SENTENCE

A. Predicate, arguments, circumstantials

Since the sentence is taken as the unit of treatment, it is necessary to adopt a formal representation of each analyzed sentence: Such a representation is important since it serves as the basis for the selection of correct word equivalents in the target language and the arrangement of those words in the proper order to form a grammatical sentence in the target language. At

the most basic level a simple and uniform representation is obtained by considering each sentence as consisting of a predicate (e.g., the verb)[3] and its arguments (e.g., the subject and object(s) of the verb), with possibly some "circumstantial" elements in addition (e.g., sentence adverbials which are not part of the subject or object(s) of the verb or of the verb phrase). The basic form of the sentence then becomes:

(6) PREDICATE + ARGUMENTS (+ CIRCUMSTANTIALS)

The order in which the terms are written in (6) does not indicate the order in which the corresponding elements actually occur in a sentence. In fact, that order is not constant within one language, but varies with the type of sentence (active, passive, interrogative, etc.); and it also varies from one language to another. (6) is independent of sentence type or language.

In (7) and (8) the parts of the sentences are identified in terms of (6):

(7)	Apparently	John	likes	the new girl
	Apparemment	Jean	aime	la nouvelle fille
	CIRC	ARG 1	PRED	ARG 2

(3) Words other than verbs may be taken as predicates (e.g., adjectives) and their dictionary entries reflect this fact, as will be explained in following paragraphs.

(8)	The new girl	is well liked
	La nouvelle fille	est bien aimée
	ARG 2	PRED

Note that in (8) the 1st argument of the predicate is not expressed although we assume that "someone" likes the new girl, and there is no circumstantial element present in the sentence. Nevertheless the abstract representation of both (7) and (8) will include PRED + ARG 1 + ARG 2. In short, we provide slots for possible arguments of the particular predicate whether or not the slots are all filled. Thus if a sentence is represented by PRED + ARG 1 + CIRC this indicates that its predicate can have only one argument (e.g., occur, exist, arrive), not that all arguments except the first have been omitted.

Recapitulating: for simplicity and uniformity, information about a sentence may be organized in the form (6) regardless of the form in which the sentence occurs in the original text or the form it takes in the translated version. Thus (6) is the first step in organizing the apparent infinite variety of sentences in a language; all sentences get the same basic representation at this level.

This basic representation is also reflected in dictionary entries where predicate words are followed by an indication of their possible arguments. This is important since restrictions on the kinds of words and phrases that can form the complement of a verb, for example, are statable in terms of its arguments, and these same restrictions recorded in the abstract representation (6) of the

sentence containing that verb provide valuable information for selecting correct equivalents in the target language. But circumstantials are not indicated in verb entries in the dictionary since they are not part of the complement of the verb, but form additional elements at the sentence level. The restrictions mentioned above ("selectional" restrictions) hold between the predicate and its arguments, not between predicate and circumstantials, hence [PREDICATE + ARGUMENTS] forms a sort of core within (6). This does not mean that there are no restrictions whatever on the kinds of circumstantials that can occur with a given predicate in a sentence, but it would be very difficult to state those restrictions given the diversity of circumstantials and the sometimes tenuous links involved. For example, to "John reads books" we might add "in the evening at home after supper under the big lamp while listening to the radio... presumably..."; but the restrictions on the arguments of read are quite precise: it is not too difficult to delimit the class of readers and the class of things that can be read.

B. Normalized structure

The abstract syntactic-semantic representation of the sentence indicated in the large rectangle in Figure 2, which is referred to in TAUM's System as the normalized structure of the sentence, is simply an elaboration of (6). Sentences may be conjoined, one sentence may be embedded within another, noun phrases may be extremely complex, circumstantials may include sentences and strings of prepositional phrases as well as adverbs, etc. But however complex a sentence may be, its basic structure will be of the form (6). And in the event that a circumstantial or one of the arguments of a predicate contains a sentence, that sentence is also represented in the basic form (6) - and so

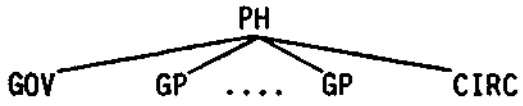
ad infinitum. In fact, each occurrence of a predicate word in a sentence gives rise to a PREDICATE + ARGUMENT component in the normalized structure of the sentence. Thus, since adjectives are taken as predicates in the system, an adjective + noun combination is represented as predicate + argument just as though it were a sentence[4] (e.g., the new book may be thought of as related to the sentence the book is new). Restrictions on the kinds of nouns that an adjective can modify are therefore treated in the same way as restrictions on the kinds of nouns that can serve as subject or object of a verb. This results in one general framework for the treatment of selectional restrictions, namely PREDICATE + ARGUMENTS; i.e., the domain of selectional restrictions is the predicate-argument relation.

The abstract syntactic-semantic representation, or normalized structure, of the sentence is at the heart of the linguistic model used by TAUM. It is pivotal in the system - the end product of analysis, containing all the information we are able to extract from the sentence, and the point of departure for transfer and generation. It is, as mentioned above, an elaboration of PREDICATE + ARGUMENTS (+ CIRCUMSTANTIALS). In the actual presentation of normalized structure the symbols used in place of the terms 'predicate' and 'argument' are GOV ("governor") and GP ("groupe prépositionnel") respectively; thus the predicate may be thought of as "governing" its arguments and these are all placed under GP for uniformity. The symbol used for 'sentence' is PH ("phrase"). The structure of the sentence at the most basic level is then represented as in (9):

---

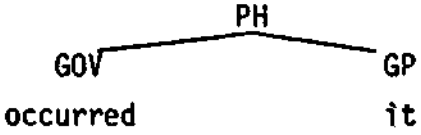
(4) The details of this representation of adjectives are given in the document on normalized structure.

(9)

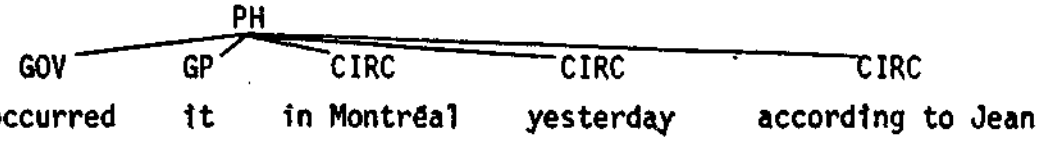


The first GP represents the first argument of the predicate, the second GP the second argument, etc. Every predicate is considered to have at least one argument, consequently GOV is always followed by at least one GP. In the case of an intransitive verb such as occur under GOV, there is one GP (the subject of occur) following GOV. In the case of a transitive verb which takes both a direct and indirect object - e.g., give - there are three GP's following GOV (subject, direct object, indirect object). The number of GP's depends on the particular predicate under GOV, but the number of CIRC's does not; there may be any number of CIRC's following the last GP, or none at all, depending on the number of circumstantial elements in the particular sentence. This is illustrated by the representations of It occurred (10) and It occurred in Montréal yesterday according to Jean (11).

(10)



(11)



Of course there will be an elaboration of GOV, GP and CIRC in normalized structure to account for the tremendous variety of predicate phrases, argument

phrases and circumstantials that can occur in sentences of the language. One could imagine under GOV a predicate phrase such as must have been being rapidly drained or an argument phrase such as the lazy old dog who usually slept under the porch and barked whenever the mailman entered the yard to deliver deliver a letter under a GP. Further elaboration of (9) is detailed in the document on the normalized structure, by Jo-Ann Stanton, March 1981.

C. Selectional restrictions

As mentioned in 5.A, selectional restrictions are used to determine, for a given predicate word occurring in a sentence, which elements of the sentence are possible arguments of the predicate. These "restrictions" are stated in terms of syntactic and semantic features assigned to words in the dictionary. For example, the verb occur takes just one argument and the dictionary entry indicates that argument must bear a feature designating an action, an event or a defect (a breakdown occurred is acceptable, whereas *a motor occurred is not; physical objects do not occur, they exist). The dictionary entry for the 3-argument verb pour (see (5), section 2-C) has selectional restrictions which indicate that HUMANS pour FLUIDS and the pouring action can be further specified by MOVEMENT prepositions such as into, onto, etc.

All this seems rather obvious, but the parser requires formal instructions to recognize the obvious. Suppose, e.g., the parser encounters the string of words cockpit left shelf. We recognize this at once as a noun phrase denoting a certain shelf in the cockpit; the parser will, in addition, interpret the string as a sentence indicating that the cockpit has departed from the shelf - unless selectional restrictions are provided to reject this interpretation. In fact, the word cockpit bears the feature COMPARTMENT and there is

no feature on the first argument position of the dictionary entry for the verb leave that would permit a "compartment" word (cockpit, cabin, compartment, bay, etc.) to be accepted as first argument of leave.

Selectional restrictions are also applied to adjectives since these are treated as predicates in our system (the noun which an adjective modifies is the first argument of the adjective). Thus the dictionary entry for abrasive indicates that it can accept as first argument a noun with the feature MATERIAL, but it is not marked for acceptance of abstract nouns.

The assignment of selectional restrictions is affected by the subject matter of the texts being analyzed. Consider, e.g., the verb deposit and the noun check. In a text dealing with banking practices check may occur as second argument (direct object) of deposit, but not in aviation maintenance manuals. The noun check exists in the latter only in the abstract action sense: maintenance personnel perform checks, but do not deposit them (in these maintenance manuals). Restrictions on adjectives likewise vary with the subject matter: we find eccentric wear patterns in the maintenance manuals, but not eccentric pilots.

In summary, selectional restrictions between predicates and their arguments are stated in the TAUM system by means of features assigned to words in the dictionary. If a noun does not bear any of the features assigned to the n^{th} argument position of a certain predicate word, that noun will be rejected as a possible n^{th} argument of the predicate; if the noun bears at least one

of those features, it is then a candidate for n^{th} argument of the predicate. And the actual assignment of features to particular words in the dictionary depends to some extent on the sublanguage under investigation.

Semantics in the current TAUM System is based on semantic features. These are essentially subcategories of nouns (and, to a lesser extent, of other categories) that reflect natural semantic categories in the "real world". Some are so widely used that they are often referred to as "universal" (ABSTRACT, CONCRETE, PHYSICAL OBJECT, ACTION, etc.) while others are more specific to certain subjects (SOLVENT might be quite useful in chemistry texts, IGNEOUS in geology texts, etc.). But the actual choice of the set of features now in use at TAUM was dictated by their usefulness in stating selectional restrictions on the predicates in the texts being analyzed.

D. Transformational relations

In addition to the relations between elements within sentences there are certain relations between sentence types that are of interest. For example, active/passive (John loves Mary / Mary is loved by John), sentences with interchange of direct and indirect objects (John gave a ring to Mary / John gave Mary a ring), etc. The most obvious relation between the corresponding sentences is that one is a paraphrase of the other. But another factor which is quite important is that the same selectional restrictions hold within each of the corresponding sentences even though the order of the words may be different and one sentence may contain certain words not found in the other.(5) Thus the class of possible 1^{st} arguments or 2^{nd} arguments of the verb love will be the

(5) This was, in fact, one of the original arguments of Chomsky (1957) and Harris (1957) in support of grammatical transformations.

same regardless of whether love is in the active or passive form. Likewise, the class of 1st, 2nd or 3rd arguments of give will be the same regardless of whether the 3rd argument (indirect object) follows the 2nd argument (direct object) and is introduced by to, or precedes the 2nd argument and is not introduced by to. Following are some examples of sentences that are transformationally related in our system.

(12) a. The filter removes foreign particles from the line.
b. Foreign particles are removed from the line by the filter.

(13) a. The hydraulic system configuration gives highest priority to the flight control boosters.
b. The hydraulic system configuration gives the flight control boosters highest priority.

(14) a. The parts removed must be cleaned.
b. The parts which are removed must be cleaned.

(15) a. the red marker
b. the marker which is red

(16) a. Select tank which is to be filled.
b. Select tank to be filled.

(17) a. To remove the cover is difficult.
b. It is difficult to remove the cover.
c. The cover is difficult to remove.

(18) a. These parts are likely to fail.
b. It is likely that these parts will fail.

(19) a. Slow down the operation.
b. Slow the operation down.

(20) a. (Something) rotates the shaft.
b. The shaft rotates.

(21) a. Presumably, this is correct.
b. This is, presumably, correct.
c. This is correct, presumably.

(22) a. Check the speed of the rotor.
b. Check the rotor speed.

(23) a. All the pumps are working.
b. The pumps are all working.

The preservation of selectional restrictions means that corresponding sentences can be assigned the same normalized structure since the order of predicate, arguments and circumstantials is fixed there anyway and selectional restrictions are stated between predicate and arguments regardless of their order of occurrence in the input sentence. It is only necessary to assign a marker to one of the related sentences (in normalized structure) to indicate that the input sentence was, e.g., passive rather than active or that the indirect object preceded the direct object, etc.; such information must be available to the transfer stage. In the analysis dictionary a verb (or other predicate word) can be given a single dictionary entry in which the relevant selectional restrictions are indicated, rather than requiring say one statement of selectional restrictions for a verb in the active form and another for a verb in the passive form, etc.

A single representation of such related sentences in normalized structure and a single statement of selectional restrictions in the dictionary is, in fact, the policy at TAUM, and it implies the incorporation of transformational relations in the linguistic model of the system.

E. Role of analysis

The role of analysis is to provide the information required by transfer and generation for finding appropriate word equivalents and generating the correct sentences in the target language. That information includes, for each sentence, the grammatical categories of the words (noun, verb, adjective, etc.)

their grammatical functions in the sentence (subject, object, etc.) and the semantic classes to which each word belongs (abstract, concrete, fluid, action, etc.). Of course, knowing the grammatical functions of all words within a sentence entails an understanding of the complex relations that hold between the words in the sentence.

In the analysis dictionary words[6] are accompanied by their grammatical categories, the semantic classes to which they belong and their syntactic complementation, including number and types of arguments where this is applicable. A word may belong to more than one category (filter is both a noun and a verb), or it may have more than one meaning within a single category (line, as a noun, may designate a geometric entity, a conductor of some sort, etc.). Once of the major tasks of the parser is to determine which of the many possible meanings is appropriate in a given context. A sentence is scanned from left to right and the parser, which incorporates a grammar of the source language,[7] makes hypotheses about the role of each word on the basis of its properties stated in the dictionary and its place in the sentence. Selectional restrictions coded in dictionary entries permit certain combinations of words to be taken as constituents in the sentence structure and reject others; as constituents are recognized, the information is stored in registers. When the entire sentence has been scanned the parser uses the contents of those registers to build a normalized structure that represents all the information it has been able to extract from the sentence. The form which this structure takes must be one which allows easy access by transfer.

---

(6) Only the base forms of words are listed in the dictionary; e.g., engine (but not engines), remove (but not removes, removed, removing). Morphological rules identify the inflected forms and convert the words to the base form along with features to indicate plural, past, present participle, etc.

(7) The parser is written in REZØ, an adaptation of Wood's augmented transition networks.