LINGUISTIC AND COMPUTATIONAL MOTIVATIONS

FOR THE

LOGOS MACHINE TRANSLATION SYSTEM

AN OVERVIEW

Bernard E. Scott

LOGOS CORPORATION

200 Valley Road Suite 400

Mt. Arlington, N. J. 07856

Preface to the 1998 Revision

This report was originally written in 1977 under contract to Control Data Corporation as part of a documentation series prepared for CDC in anticipation of a contemplated investment in Logos Corporation. The aim of the document was to explain to CDC management and evaluators the sort of thinking that had gone into the design of the Logos system right from the inception of work in 1970. Although addressed to an unschooled audience, the document continues to represent the most complete description available of the motivations

and principles underlying the design and implementation of the Logos system. For that reason, its distribution to professional linguists, computational linguists, and workers in cognitive science who wish to know something of the Logos system is thought justified.

The original 1977 manuscript has been edited only very slightly, in most cases simply to correct lapses of one kind or another. Nothing of substance has been changed regarding the system itself or its underlying principles, except in the few instances where noted explicitly. Obviously, a great deal of work has been accomplished on the system in the intervening twenty years, which cannot be reported on here. And much of what does appear may possibly strike the present reader as simplistic and even primitive. All that notwithstanding, the translation model conceived in 1970 and implemented over the following years has held up exceedingly well. No one argues that the Logos system today is not state of the art, at least in terms of the quality of translations it produces for multinational customers in 12 or so nations. That an unbroken line of thinking has guided the design and development of this system for almost thirty years may be the most remarkable thing to be said about the system and this twenty-year old description of it.

Bernard (Bud) Scott

January 1998

I. LINGUISTIC MOTIVATIONS

A. Definition of the Linguistic Problem confronting Machine Translation (MT).

Machine translation may be succinctly defined as the mapping of one language into another by electronic means. In practical terms, such mapping entails the manipulation of a meaningful string of words of a given natural language, formulated in accordance with the grammar of that language, such that a semantically equivalent string of words is produced in another language, in accordance with the grammar of this new language.[1]

Manipulation of the input word-string to produce a target language equivalent is of the following kinds (illustrated with English as source language and French as target language):

1. Lexical substitution:

(1) typewriter ----> machine à écrire

2. Inflections and syntactic re-ordering:

(2) the blue pencils ----> les crayons bleus

3. Syntactic equivalencing:

(3) children are playing here ----> des enfants jouent ici

4. Syntactic transformation:

(4) the man should be told ----> on doit informer l'homme.

5. Stylistic transformation:

(5) X provides an indication of Y ----> X indique Y

Manipulations 1 through 5 in themselves pose no inherent difficulty for an MT system. What constrains an MT system is its ability to analyze which manipulation is called for and where, so as to synthesize the target language equivalent correctly. And what constrains the ability to analyze a source language sentence is the prevalence of ambiguity in the source language.[2]

We define ambiguity as a linguistic situation capable of more than one interpretation. The MT system encounters such ambiguity at any of six linguistic levels in the source language sentences. These levels are:

(i) lexical-syntactic, or the parts of speech as they appear in the dictionary.

(ii) sentential-syntactic, or scope and dependencies of parts of speech as ordered in a sentence.

(iii) lexical-semantic, or meanings of words as found in a dictionary.

(iv) simple sentential-semantic, or meaning of words as a function of simple contexts.

(v) complex sentential-semantic, or the meaning of words as a function of complex contexts.

(vi) extra-sentential, where resolution of anaphora, ellipses, etc. depend upon information in previous sentences.

Each linguistic level is liable to ambiguity peculiar to that level alone. Ambiguity at any one level is typically resolved by referring to information available at the next higher linguistic level. For example:

a. lexical-syntactic ambiguity is lexical ambiguity as to part-of-speech. For example, the word "check" is found in the dictionary both as a noun and as a verb. The MT system encounters its first ambiguity when it looks up the word "check" and finds a part-of-speech selection must be made. At this stage, the MT system does not know which part-of-speech to select. The process of resolving this lexical-syntactic ambiguity entails referring to the next higher linguistic level, namely, the sentential-syntactic level. By analyzing how the word "check" functions in the sentence, the system presumably resolves the ambiguity.

b. sentential-syntactic ambiguity is ambiguity as to the syntactic function a word has in a given sentence. A typical ambiguity of this kind relates to the question of governance. For example, in the syntactic string (a prepositional phrase, where N represents noun):

(6) to N1 of N2 and N3

it is not clear whether N3 is governed by the preposition "to" or the preposition "of" (or possibly by neither preposition). Once again, the matter is resolved only by referring to the next higher linguistic level, viz. the lexical-semantic level, that is, by examining semantic values for the nouns N1, N2 and N3, as given in the dictionary.

To illustrate, consider the following prepositional phrases based on (6):

(7) to citizens of Rome and environs

(8) to citizens of Rome and friends

In (7) the semantic kinship of N2 ("Rome") and N3 ("environs") causes the preposition "of" to govern N3 as it does N2 (unambiguously). In (8) the semantic kinship of N1 ("citizens") and N3 ("friends") causes the preposition "to" to govern N3 as it does N1 (unambiguously).

c. Lexical-semantic ambiguity arises because a given dictionary entry has more than one semantic meaning to choose from within a single part-of-speech. The

transitive verb "check", for example, may mean "examine" or "consult". While this verb conveniently bears both meanings in English, separate verbs must be used for each meaning in most target languages. The MT system, therefore, must resolve this ambiguity. It does so by referring to the next higher linguistic level, namely, simple sentential-semantic. That is, the particular nuance of the verb "check" is understood by examining the semantic environment, usually its immediate semantic environment.

The following two sentences will illustrate:

(9)(i) Check the weather bureau before departing.

(9)(ii) Check the engine compression before departing.

In (9)(i) the meaning of "check" is "consult". In (9)(ii) "check" means "examine". This is established by the semantic properties of the direct object of the verb.

d. Simple sentential-semantic ambiguity arises when some word or group of words in a sentence is capable of meanings not usually given for that word in the dictionary. This happens because of the influence on that word's meaning by other elements in the sentence. The following sentences will illustrate:

(10) (i) Check the newspaper for dates.

(10)(ii) Check the newspaper for errors.

In (10)(i) the preposition "for" acquires the meaning "for information concerning" and must be so rendered in more than one target language. (In Vietnamese, for example, "for" in (10)(i) is rendered dê biêt, "in order to know"). In (10)(ii) "for" has the meaning of "for the presence of". It should also be noted that the verb "check" translates to "consult" in (10)(i) and "examine" in (10)(ii) and that the proper selection of the verb's nuance depends not merely on the object of the verb, as in (9)(i) and (9)(ii), but on the entire sentence. The same is true for the preposition "for".

e. Complex sentential-semantic ambiguities are of two types:

(a) ambiguity as to antecedence of pronouns or other anaphora.

(b) ambiguity as to the meaning of a word, particularly true of common nouns, where the particular meaning is determinable only by information available outside the sentence. For example:

(11) He ate four ears of corn. The ears were very large.

In both (a) and (b), resolution depends on extra-sentential information which an MT system must "remember" or carry over into its analysis of the current sentence.

From the foregoing analysis it should have become apparent that the fundamental problem for MT technology is that of simulating human linguistic processes of a fairly high order. Is this possible, and if so, how is it done?

To answer this question in light of Logos machine translation technology, a discussion of the linguistic principles underlying Logos technology is appropriate.

B. Linguistic Principles Underlying the Logos MT Solution

1. General Observations

Noam Chomsky, in 1957, voiced a germinal question regarding language, the answer to which underlies much of modern linguistic investigation. He asked: what mental process in a child enables it to utter grammatical sentences it has never heard? Put otherwise, what linguistic process takes place in a child's mind that enables it, after a finite exposure to language, to generate sentences of infinite variety?

The question is germane to foreign language study as well. We ask: how is it that a student of a foreign language at some point begins to utter sentences in that language which he has never heard? Clearly, some linguistic process is involved beyond that of imitation. Has this principle been identified?

In the discussion which follows, we present an analysis of certain linguistic processes believed to be involved in human learning of a second language and show how these processes have influenced the character of the Logos system.

2. Linguistic Process Involved in Translation

In learning French, an English-speaking student at some point is taught to express in French the following sort of English sentence:

(11)(i) John wants Mary to study music.

(11)(ii) Jean veux que Marie s'applique à la musique.

In the next instant when the student attempts to apply what he has just learned about the construction of French sentences to another English sentence similar in structure to (11)(i), such as:

(12)(i) John asks Marie to study music.

he is told, in this case, he must use a different construction:

(12)(ii) Jean demande à Marie qu'elle s'applique à la musique.

or

(12)(iii) Jean demande à Marie de s'appliquer à la musique.

Now, suddenly, the student is confronted with a problem: constructions, which appear identical in English, require different constructions in French. Faced with this ambiguity (by our definition), the typical student hesitates to render such constructions into French until he "knows French better". The question is, what does he need to learn?

The situation is no better for the MT system. In its initial effort to deal with English sentences (11)(i) and (12)(i), the MT system "perceives" a common syntactic construction in the English (N = Noun, V = Verb):

(13) N1 V1 N2 to V2 N 3

From such a syntactic string, how shall a machine be further programmed to synthesize now one French construction, now another?

Obviously, it is the verb V1 that differentiates (11))i) and (12)(i). Certain verbs in French require certain constructions, other verbs other constructions. Are these constructions arbitrary conventions of a given language or are they a function of something independent of convention (such as the verb's meaning)? And must the student, to master French, study each verb of the French language for the construction it takes, verb-by-verb, or is there some underlying principle the gifted student intuitively grasps and makes use of, so that he begins to express himself freely in the foreign language after only a short time? Must the MT system, in its turn, be instructed as to the behavior of each verb, or is there a more elegant way to handle the problem of verbs and their behavior?

To answer these questions, we need to observe an important property of language having to do with its conceptual basis. First, we note that the English verbs used in (11)(i) and (12)(i), namely, "want" and "ask", can also take optional English constructions that parallel their French counterpart, in (11)(ii) and (12)(ii) respectively:

(11)(iii) John wants that Mary study music.

(12)(iv) John asks of Mary that she study music.

From this we might infer that "want" and "vouloir" in (11)(iii) and (11)(ii), for instance, take parallel constructions, in part at least, because of the common semantic value which these French and English verbs share. A parallel situation exists with the verb pair "ask" and "demander" in (12)(iv) and (12)(ii). This suggests that syntactic constructions are not purely accidents of a given language but are shaped, in part at least, by the semantic values of the verb as such, independent of a given language and its grammar. This inference is reinforced by several other observations. One, the verbs in the English constructions (11)(iii) and (12)(iv), "want" and " ask", are not convertible without a shift of meaning.[4] This again implies the existence of a strong connection between semantic values and syntactic behavior in the case of verbs. Two, when we substitute other verbs whose syntactic behavior fits the patterns of (11)(iii) and (12)(iv), we discover an important fact: verbs that behave alike syntactically look alike semantically (roughly speaking). For example, the verbs "request", "entreat", "beg" belong to the same semantic family as "ask" and exhibit the same syntactic behavior as "ask" in (12)(i) and (12)(iv). This is generally true of their French counterparts as well.[5]

The conclusion to be drawn from the foregoing is that syntax and semantics are not separate and unrelated properties of language, as Chomsky's dichotomy between deep structure and surface structure implies, but are rather intrinsic aspects of the same property, designated as a semantico-syntactic property (See Appendix A, "Is Syntax Ever Independent of Semantics?").

The fundamental connection between syntax and semantics becomes clearer when we consider verbs in the light of the degree of transitivity which they exhibit (see Figure 1). The degree of transitivity is a function of the verb's meaning (semantic value), while at the same time, the degree of transitivity that the verb has, has a pronounced effect on the syntax of the sentence.[6] A verb is not transitive because it takes an object; it takes an object because it is transitive, because the verb semantically implies an object, and, in fact, cannot stand without it, because the

verb's meaning entails a movement from the subject across to something, etc. Both the movement itself and the nature of the object toward which this movement is directed varies accordingly as the main verb varies in this matter of its degree of transitivity, which we hold to be a semantic, or more properly, a semantico-syntactic property.

We may think of syntax and semantics vis à vis verbs as the extreme terminals of a semantico-syntactic verb tree (see Figure 2).[7] At the semantic extreme of this tree are the leaves, i.e., all the verbs of a given language, fully individualized. At

the syntactic extreme of this tree is the tree trunk, which embraces all verbs in the single syntactic symbol V. In between are the branches, which represent different semantico-syntactic properties. As we move up the trunk out onto one or another of the branches, sub-branches, etc., each designating a set of verbs characterized by certain distinct syntactic properties, we find the semantic shape of these verbs also begins to emerge. We find, in effect, that the syntactic property and the semantic shape are interrelated.

To illustrate the interconnection between syntax and semantics as it pertains to semantico-syntactic trees, let us process a string of words representing a simple sentence (where W = word).

(14) W1 W2 W3 W4 W5

Now, let us assume each word in (14) is to be found somewhere on a semantico-syntactic tree, and that there is one such tree for all verbs, one for all nouns, and so on. Next, let us assume we climb the semantico-syntactic tree for each word in (14), starting at the bottom of the tree, at the extreme syntactic end. As we do so, we uncover the primitive syntactic value for each:

(15) N1 V1 N2 prep N3

W2, we perceive, is a verb (V1). Since the verb dictates many (if not all) of the relationships the other elements in the sentence will have to each other, we immediately focus on the semantico-syntactic tree for V1.

As we move up the tree, we find that the trunk forks and that V1 takes us out onto the branch reserved for verbs that take direct objects. We know very little about V1 at this stage, but a modicum of semantic shape has begun to emerge from the meaningless W2: V1 represents an action that N1 performs vis à vis an object, presumably N2

We move further along this limb and discover that V1 enters another branch reserved for verbs which take two objects, designated split transitive verbs.[8] Split transitive verbs are verbs which have a direct and an indirect object.

We now know considerably more about the semantic character of V1. We

know, for one, what V1 could not be, semantically speaking. We also know V1 must have the general semantic sense of "impart", a sense common to all split transitive verbs:

(16) N1 deals N2 to N3.

N1 relates N2 to N3.

N1 supplies N2 to N3.

Moving further along the semantico-syntactic tree for verbs, we find that V1 enters yet another sub-branch reserved for split transitive verbs that have still another important syntactic property, namely, that they govern the preposition "with" in such a way that in (14) the object of the preposition "with", N3, is the true direct object of the verb V1, and that N2, the apparent object of V1, is in reality the indirect object of V1.

V1 is an "impart type" verb, therefore, whose syntactic behavior represents a minor variant of the more usual pattern for split transitive verbs.

(17) N1 V1(impart-type) N2 with N3 -->...

...--> N1 V1(impart-type) N3 to N2

Let us say also that the preposition in (15) happens to be "with".

We now found ourselves on a small sub-branch of the verb tree. The number of impart-type verbs supported by this sub-branch is small. The semantic shape of the verb is fairly clear; of the verbs in list (16), only the verb "supply" qualifies for this sub-branch:

(18) N1 V1(supply-type) N2 with N3

In the foregoing, we have seen how from the mere accumulation of a verb's syntactic properties, the verb's semantic value has all but emerged. There can be little doubt that syntax and semantics are organically related.[9,10]

We need now to pick up again with the beginning student of French. Were he given an English sentence to translate, such as:

(22) John provided the students with books.

his first impulse will mostly likely be to render "with" as "avec", (the principal lexical transfer) much as early-generation MT systems did and low-end MT systems still do. The teacher shows the student the correct translation:

(23) Jean a fourni des livres aux élèves.

By way of helping the student appreciate the differences between English and French, the teacher proceeds to illustrate the variety of ways in which the preposition "with" is to be rendered in French:

(24) John sees a man with a dog. (with = avec)

(25) John blocks the flow with a valve. (with = au moyen de)

(26) John acquaints the man with the facts. (with = des)[11]

(26) John does not mix with the students. (with = à)[12]

However, the teacher does not explain to the student when or why each usage becomes appropriate, except by way of illustration on a sentence-by-sentence basis.

The student must figure out the why by himself as best he can, through the mysterious linguistic process called "learning" a language.

Let us now see what we can fathom of this process in light of the discussion thus far.

If the student is reasonably gifted, he intuitively grasps that the preposition "with" loses its simple lexical meaning as "avec" in (23), (25), (26) and (27) because the verb in each of these sentences makes a prior claim on it. In (24), by the same token, "with" can be rendered as "avec" just because this verb does not make a claim to it. Without necessarily formulating it as such in the mind, the student learns to be sensitive to prepositions and to the fact that they might belong, semantico-syntactically speaking, to a verb and as a result take on an unexpected meaning.

Next, in a largely unconscious way, a mental sorting takes place as to why one verb causes a preposition to function one way, with one meaning, and another verb causes it to function another way, with another meaning. Now, an especially gifted student at this point, without necessarily realizing it, finds that as he studies the constructions (22) through (27), the changes which the preposition "with" undergoes in the French in each case seems entirely appropriate. He does not articulate it, but the student is aware that something in the nature of the verb seems to call for what happens to the preposition. For example, the verb "block" calls for explanation as to how, with what instrument, by what means this action is performed. It is difficult to think of the action "to block" apart from an instrument. When the student encounters this verb, an expectation is unconsciously set up for the means. Then when the student encounters the preposition, his mind grasps it as supplying what was deficient in the verb itself. In fact, the meaning of "with" as "by means of" rushes to mind as if it were filling a semantic vacuum. Conversely, in (24) the preposition "with" is taken as a noun

complement just because the verb "see" does not create an expectation of means, does not suffer a semantic deficiency (except as to its object, which is true of all transitive verbs), does not need the preposition to complete its thought.

The situation is the same for (22), (25), (26) and (27). For example, the verb "acquaint" in (26) sets up the expectation as to about what. The prepositional phrase "with N3" satisfies the "about what" expectation, and in the French, the "with" is rendered by "de" about.[13]

Simultaneous with the student's unconscious grasp of the inter-connection of syntax and semantics in the case of the verbs and prepositions we have been discussing, the gifted student intuitively grasps another principle: the interconnection between a verb and its preposition holds true not only for this verb but for all other verbs that are like this verb, that is, for all verbs possessing the same semantico-syntactic property of needing the preposition to complete their sense. For example, having seen the verb "block" in (25), the student has no trouble recognizing that other verbs such as:

(28) compensate

cap

balance

control

choke

all relate to the preposition "with" in the same way, and therefore all take au moyen de as the translation of the preposition. By the same mental operation, he recognizes other "with"-oriented verbs do not belong to (28), as, for example:

(29) align

synchronize

coordinate

In (29) the verbs impart to "with" the meaning "with respect to". He sees that the verb "align" in (29) behaves like verbs in (28) when the object of the preposition "with" is an "instrument type" noun. For example, "align the head with a wrench"; otherwise, "align the head with the sides of the ... etc".

How does he know these things? Let us see if we cannot account more formally for the linguistic skills such a gifted student exhibits.

Using verbs as our focal point, we delineate three planes on which linguistic mastery develops. First and most fundamentally, there is the immediate analytic knowledge as to the verb's behavior, how it relates to other elements in the sentence, which knowledge the student intuits directly from the semantic nature of the verb itself. That is to say, he grasps the verb's syntactic properties from its semantic character, as something implicit. Although the student does not think of it as such, his mind penetrates the verb at the point where semantics and syntax intersect, which point serves as the basis for the verb's place on the semantico-syntactic code. Although the mind does not articulate these codes, there is reason to think that these codes do indeed articulate what the mind is dealing with in linguistic operations of the sort.[15]

To illustrate this initial plane in linguistic mastery - which we have defined as an unmediated analytic knowledge of a word's behavior gleaned from the word itself - let us observe the linguistic process that occurs when the student encounters the verb "protect". This verb announces loud and clear its need for complementation as to what the protection shall be with respect to. This expectation of complementation is implanted in the student's mind as he reads or hears the verb, such that should the student encounter a sentence such as:

(30) A scarf protects his face X the wind.

there is no hesitation in his mind as to both the syntactic function and the semantic import of X. X is a preposition with the general meaning of "with respect to".

This characterization of X, moreover, will be true regardless of which language the sentence (30) might be rendered in, so it is something which is true (and therefore knowable) independently of conventions of any kind. However, the student will not know on this basis a single preposition in any given language that actually fulfills the semantico-syntactic functions of X. This is an empirical kind of knowledge, an acquired knowledge that indeed must be gained from the conventions of each language. This we term the second plane of linguistic mastery, the plane on which exposure to usage takes place, with such activities as memorizing word lists, use of dictionaries, and grammatical exercises of every kind. It is a plane of linguistic learning that pedagogy ordinarily treats as the only plane. That it is not the only plane, however, can be inferred from the question for which pedagogy and linguistics have as yet found no answer - namely, what enables the gifted student to very rapidly advance beyond rote learning and mere imitation to begin expressing himself freely in a new language? We answer that the reason is to be found in the planes of knowledge which surround this plane of learning by imitation.

As for these conventions as they bear on X, the student acquires from the convention of Russian usage that X is rendered by "OT" (from) and from the convention of French usage that X is rendered by "contre" (against) or "de" (from). Not untypically, he finds that English convention also allows both "from" and "against" equally. That two such disparate prepositions should serve the same semantico-syntactic function when used with the verb "protect" illustrates, we think, both the freedom convention enjoys and the fact that the contingent character of X is possible just because of the semantico-syntactic necessity imposed on X by the verb, which insures its meaning no matter what convention does with X.

It seems reasonable to assume that sensitivity and prowess on the first linguistic plane re-enforces the ability of the student to learn the conventions of grammar for a given language, for although the specifics of X are not dictated by the verb, nevertheless, the appreciation of X's semantico-syntactic property must surely influence that selection and most certainly help the student to remember it. But it is not this that enables the student to branch out creatively in a language. This final stage of mastery, though it depends on the first and second planes, takes place by the developments which occur in the student's mind on yet another plane.

The knowledge developed on this third plane is analogical knowledge, knowledge gained through perceiving the proportional similarities between things, which differ materially. It is by this knowledge that the student's grasp of a language begins to spread like wildfire. For example, the student, having learned what he knows about the verb "protect" and the conventions which implement its prepositional complement, when he thereafter encounters a verb like "insulate", he perceives first that this verb takes complementation as to what the insulation shall be with respect to. (This is the analytic knowledge developed on the first plane.) Next, and most importantly to the learning process, he then sees that in respect of its semantico-syntactic property, "insulate" is very much like "protect", and that, therefore, the conventions which apply to "protect" should also apply to "insulate" (analogical knowledge). In the next instant, the student is using the verb "insulate" with correct complementation even though he has never had the verb explained to him before.

This then is what goes on in the gifted student's mind. When he encounters a new verb that he has never seen used, he first tastes its properties, so to speak; then he casts about in his mind until he finds an analogy for it among those verbs with which he is familiar, and finding one he proceeds to fit the new verb into the conventions that apply to the known verb.[16] Without doing so explicitly or consciously, the gifted student operates with semantico-syntactic properties such as we have been describing, and by such stepping stones picks his way surefooted among the multiplicity of verbs in a language, using the power of analogy to reduce the multiplicity to a finite set of linguistic situations. Conversely, having mastered these few score situations, he expands his

command over the whole of language in all its endless variations, all by the power of analogy.

The gifted student thus makes remarkable progress, seemingly able to handle himself in a new language with a swiftness and ease that baffles the less gifted. As a matter of pedagogy, it seems that what the gifted student comes to by virtue of his intelligence, the slower student could be taught to see. The semantico-syntactic behavior of words, and the grouping of words on the basis of analogous behavior, are aspects of language that the teacher could point out. The student, it seems to us, would grasp such aids because they conform to the linguistic process natural to the mind.

An indirect proof that this is indeed the way the mind works is afforded by the old vaudeville repartee that begins with the interlocutor asking:

(31) Would you hit a woman with a baby in her arms?

to which is given the reply:

(32) No. I'd hit her with a brick.

Why is the line funny? It amuses us because of its paradoxical quality of having a certain plausibleness in the face of its manifest absurdity. The plausibleness stems from the expectation that the verb "hit" sets up in our minds. We do indeed listen for the instrument after hearing the preposition "with", but then immediately shift semantico-syntactic gears, so to speak, on hearing "baby", which naturally complements "woman" and makes of the prepositional phrase a noun complement. It is funny because it is a syntactic pun, and like any pun, depends for its humor on the plausibleness of the double entendre. Puns are painful when the double entendre is forced and unlikely. The best puns are those which catch us red-handed, as in (32); then the laughter is a kind of admission, an acknowledgment of our own mental operations.

What does all this mean practically? It means that a student does not have to learn 10,000 verbs and their constructions in order to master French. Instead, he learns perhaps 30 or 50 constructions with a representative verb for each. Then when he encounters a new verb, he mentally compares the unfamiliar verb to a familiar verb that is like it. Using the verb he knows - whose semantico-syntactic properties he knows - he substitutes in its place the new verb that resembles it and proceeds to express himself in tried and true form, as if this verb were one he had thoroughly mastered.[17] He makes errors, to be sure, but through trial and error, he refines his knowledge and his instinct for proceeding in this fashion, learning to detect subtler semantico-syntactic features in the linguistic landscape. Such a student makes remarkably rapid progress, to everyone's amazement. Should he study a second foreign language, his skills become sharper still. Everyone, indeed, experiences greater ease learning the second and third

foreign language. But of what does this increased skill consist if not precisely in what we have described?

What does this mean practically for MT?

It suggests that the system ought to be taught to deal with language in the way the gifted student does, by perceiving words at the level in which the word's semantic properties and syntactic properties intersect. The entries in the System's dictionary, therefore, should be encoded according to these properties. The entry for the verb "protect" should have codes that identify its complementation. Thus, these codes would make available to a machine an articulation of the analytic knowledge about words, which the student intuits in a largely unconscious way. This in sum is the way the Logos system was built.

As we have seen by the example of the Logos semantico-syntactic tree for verbs, the codes used to identify the semantico-syntactic properties of a verb in the Logos system's dictionary are generic codes, meaning that the verb "insulate" will be encoded for the most part with the same codes as "protect". Thus, these semantico-syntactic codes also incorporate the analogical knowledge which is so important to the student's linguistic processes.

The computer has no mind, but it has a memory, which is to say that an input string can be processed through the computer's memory. Into the memory of the Logos system are stored linguistic rules that consist of patterns of these semantico-syntactic codes. When one of these rules is found to apply to a portion of the input string, the rule causes various manipulations to be performed on the input string, all in accordance with the conventions appropriate to the target language grammar.

Since these linguistic rules are based on patterns of semantico-syntactic codes, they are in effect, generic patterns. As such, a relatively small number of rules is able to deal with endless varieties of construction, doing so in a way that is sensitive to the semantic import of any given element as it affects syntactic re-ordering, lexical choice or stylistic transformation necessary for achieving a good translation. In short, the generic patterns of semantico-syntactic codes, supplemented as needed with unique codes for specific words, enable the System to resolve ambiguity at both the syntactic and semantic level, and to this extent, provide the basis for high quality translation.

In conclusion, it can be reasonably argued that the student's own linguistic operations also entail the grouping together as a class all words whose semantico-syntactic properties are the same, so that on encountering a new word (a new input, as it were), he first savors the word for its semantico-syntactic property (akin to a dictionary lookup to retrieve the word's codes), and then identifies it not with some other verb like it, but with the class of verbs which share the same property. It is as if the analytic plane, receiving feedback from the

analogical plane, refines its knowledge and forms classes of verbs, such that thereafter the function of the analogical plane is to subsume a new word under the correct class. Thus it would seem that what the mind deals with when it deals with the problems of syntax and with lexical choice and other matters of translation are not specific words but, at times at least, abstract entities which if they were to be labeled, would take labels that resemble the semantico-syntactic codes of the Logos system.

C. Application of Semantico-Syntactics in the Logos System

1. Recapitulation

In Section A of this chapter, we said that what limited an MT system in the final analysis was its ability to simulate human cognition of the kind involved in translation. More specifically, an MT system succeeded or failed depending on its ability to resolve ambiguities, i.e., linguistic situations capable of more than one interpretation. We saw that to resolve ambiguities at any level, the mind no less than the machine had to refer to a higher linguistic level, and that language abounded in ambiguities such that, in some cases, only by a semantic grasp of the sentence as a whole could the matter be resolved. Thus, the machine had to simulate human linguistic processes at a fairly high level. The question was raised: how was this possible?

In Section B, we described some of the linguistic operations which the mind appears to perform in resolving ambiguities that arise in translation, where an English construction sometimes takes now one French construction, now another. We saw that ambiguities of this kind were resolved at the semantic level by considering semantic properties of the verb. We then asked a practical question, touching on the very nature of the linguistic process, namely, does the student have to address each verb of the language as an independent case? This question relates to the larger question Chomsky raised about how languages are learned: of what precisely does the linguistic skill consist that enables a student to advance beyond rote learning to express himself freely in a language. We answered this question by describing a process wherein the mind deals with language at the point where syntax and semantics intersect. We suggested that the gifted student senses the semantico-syntactic properties of words and by the use of analogy with respect to such properties is able very rapidly to extend the knowledge he is acquiring as to the linguistic conventions of a given language over that entire language after only limited study.

In the present section, we will illustrate how a semantico-syntactic representation--the description of language at the point where semantics and syntax intersect--solves many difficult ambiguity problems in MT, and in so doing serves as the proper linguistic basis for an MT System. We will suggest,

furthermore, that this is the only practical method for programming a machine where the object is to simulate human linguistic processes.

2. The Elusive Goal of Semantically Sensitive MT.

Machine translation technology, after getting off to a heady start in the late 50's and early 60's, arrived at a dead-end by the late '60's when it was realized that syntactic parsing of sentences in and of itself was an insufficient basis for MT, for the simple but compelling reason that ambiguities in the syntactic string could not be resolved without recourse to semantic data. Take, for example, the following syntactic string:

(33) N1 prep N2 V'ed by N3

In (33), the participial phrase "V'ed by N3" can complement either N1 or N2. Seemingly, only an appeal to the actual meaning of the noun phrase as a whole can resolve the ambiguity in (33).

Early MT efforts (of which there were at least 100 projects worldwide) deliberately separated syntax and semantics as distinct and independent problems with syntax always being considered as the first order of business. Motives for this precedence to syntax undoubtedly relate to the close connection between mathematical logic and linguistic modeling in this early period and the fact that this connection arose out of structural linguistics. The appeal of syntax is that it allows language to be represented by a manageably small, univocal set of symbols, manipulatable in a machine environment not too unlike logical expressions, even though the kinds of operations involved were somewhat different. No such scheme presented itself for semantic representation. These early efforts focused, therefore, on syntax alone, on developing schemes whereby a machine could syntactically parse a source-language sentence. It was believed that once a firm syntactic foundation was established, a semantic superstructure would follow. But this semantic superstructure was never forthcoming, and as it turned out, syntax bereft of semantics proved at best a weak foundation.

For example, Naomi Sager and her co-workers at New York University Courant Institute have been at work for a number of years on a linguistic string parser, a very respectable effort, but one which seems stalled as much as ever by the problem posed in (33).

Sager poses the problem in the form of a noun phrase:

(34) Changes in cells produced by digitalis.

How shall a linguistic string parser be instructed to recognize whether "produced by digitalis" complements "changes" or "cells"? Acknowledging it as a still unsolved problem, Sager points hopefully to one suggested solution: if rules were devised governing permissible word-association within a subfield of knowledge, it should be possible by such rules to determine that "digitalis" can be associated with the production of "changes" but not of "cells".[18]

A system dependent on pragmatic knowledge, e.g. on all the associations and combinations of words permissible within a subfield of knowledge, is feasible only in the narrowest of bandwidths. For a field as broad as pharmacology, the number of rules needed to capture permissible word-associations would be truly immense and gaps would be inevitable. As a solution to the disambiguation problem posed by phrases like (33), therefore, the word-association concept, even if theoretically feasible, would seem to make the goal more remote than ever.

Semantic (as opposed to pragmatic) networks might appear to offer a solution. Semantic networks have indeed been elaborated in recent years (1994) that show relationships between words, but these are usually basic relationships of a defining nature. Theoretically, a network aimed narrowly at pharmacology could certainly relate digitalis to medicine, medicine to cures, and cures to change. But it could also be expected to relate medicine to body and body to cells, which likely would leave us with the ambiguity unresolved.

To our way of thinking, these approaches suffer from two basic misconceptions. First, and most fundamental, is the assumption that for a machine to resolve (33), it must "know" what a specialist knows, namely, whether "digitalis" in fact produces "changes" or "cells". That this assumption is mistaken can be seen from the fact that it takes no specialized knowledge for a reader to know that the participial phrase "produced by digitalis" modifies the noun "change". All it takes is a modicum of grammatical sense, as we shall demonstrate. This should be sufficient basis for a machine as well. This has in fact has been informally demonstrated on numerous occasions when we show prospective candidates (for employment) a variant of (33) with the symbol X substituted for "digitalis". The overwhelming majority of these candidates parse the phrase as a pharmacological specialist would have done.

The second misconception is related to the first. Semantics as an object of study has been artificially severed from syntax, a fact which we believe accounts for the difficulty researchers are having in getting a handle on semantics. We suggest that syntax itself is the handle, that just as semantic information helps us to cope with problems of syntax, so syntactic information helps cope with problems of semantics. This claim is strengthened by the fact that neither syntax nor semantics by itself is able to resolve the ambiguity in (33). On the other hand, when this ambiguity is viewed semantico-syntactically, it simply dissolves, as we shall show. Problems both of syntax and semantics, when these are treated in

isolation, tend to disappear when these are brought together, particularly at the point where syntax and semantics intersect. It is our view, moreover, that the mind, in processing language, rarely operates in a purely semantic or in a purely syntactic way, but operates rather at this semantico-syntactic level, as we shall attempt to illustrate in what follows. This finding in turn motivated the semantico-syntactic orientation of the Logos MT system.

The semantico-syntactic operations of the human mind that we discuss below represent a kind of linguistic shorthand which reduces the multiplicity and infinite variety of language to something which a young child masters quite naturally. To the extent that this "shorthand" is understood, a machine can just as easily be instructed to simulate these semantico-syntactic operations, certainly to a degree sufficient to produce good quality translation.[19]

3. Semantico-Syntactic Solutions to Semantic Ambiguities in MT

The problem posed by the phrase "changes in cells produced by digitalis" is a machine-oriented problem. The mind has no difficulty recognizing that the complement "produced by digitalis" complements "changes". But how shall an MT system decide the matter if all it "perceives" is the syntactic string:

(38) N1 in N2 V'ed by N3.

A solution by word association involves semantically encoding N1, N2 and N3 and the formulation of pragmatic rules allowing, in certain contexts, an association between N1 and N3, and disallowing one between N2 and N3. As we have attempted to argue, this is not feasible in all but the narrowest applications.

We find we can solve this problem easily enough if we instead consider N1 and N2 in light, not of N3, but of the verb "produce". As usual, the verbal element holds the key to the relationships among all the nouns in (38). The matter will be clearer if we substitute two other verbs for "produced" i (38):

(39) changes in cells effected by digitalis

(40) changes in cells affected by digitalis

We see at once that the verb "effect" by its very nature can only have "changes" for its referent. On the other hand, we see that the verb "affect", while it could conceivably complement either "changes" or "cells", in fact complements the latter. It does so by an inexorable law (for lightly inflected languages like English at least) that says modifiers shall modify the element that is closest to them unless there is a compelling reason to do otherwise. No compelling reason can be found for this in (40) so that by default the referent is the closest NP.

The compelling reason found in (39) resides in the fact that the verb "effect" is the kind of verb that requires a process noun (noun deverbal) for an object. When the mind encounters verbs of this classification, it seeks out and expects to find a process noun somewhere. "Changes" is a process noun; "cells" is not. The mind therefore leaps over "cells" back to "change" in an operation as natural and unconscious as it is inexorable, in view of the nature of this verb.

We have to assume that the mind does this because its operation here is guided by an immediate, intuitive grasp of the semantico-syntactic property of the participle "produced." If, in (38) that participle were coded as a "preprocess verb", meaning that it's object could be a process, and if N1 were coded as a process noun,

we discover that the very simple rule (41) is all that is needed for resolving the ambiguity in (38).

(41) N1(process-type) in N2 V'ed (preprocess-type) by N3

Rule (41) has merely to test (38) for the presence of the semantico-syntactic collocation. Upon finding these properties in combination, the rule effects the correct complementation. Such a rule, by the same token, would fail to detect the semantico-syntactic combination in (38) if the V'ed were "affected", since "affect" is not a preprocessl-type verb.

To get back to the original verb "produce", this is also a preprocess verb (though not exclusively so) and therefore has the same basic classification as "effect". As far as this linguistic rule is concerned, then, it applies to any process noun in combination with any preverbal. The rule thus has great generality. (To be sure, one can always create a context where "produce" relates to N2 and not N1 in (38), but almost always these examples are forced and potentially ambiguous. The author of such an expression would doubtless introduce surface structure clues to avoid the potential for ambiguity here.)

4. The Role of Verbal Elements

It will be apparent to the reader perhaps that much of our discussion up to now has centered on verbs. This is not by accident. Verbs are the key to almost every linguistic situation, though sometimes the verbal key is well hidden.

Broadly speaking, verbs are what shape utterances. They often also bear the essential point or meaning of an utterance. Without context, is neither interesting nor meaningful to posit a subject by itself, no matter what it is. To make a meaningful utterance we must predicate something of it. When we do so, we find that we use verbs or verbal elements, for verbs and verbal elements are what express relationships. Conversely, we can say that any relationship is verbal.

Conventional grammar recognizes this fact when it calls the verb the "predicate". The adjective in "John is smart" is called a predicate adjective.

There are other verbal elements not always recognized as such, however. For example, the classical ambiguity:

(42) flying airplanes can be dangerous

is ambiguous only because the verbal elements which are implied are not explicit. If they are, the ambiguity dissolves:

(43) a) The act of flying airplanes can be dangerous.

b) The proximity of flying airplanes can be dangerous.

What distinguishes the nouns "act" and "proximity" is precisely the verbal bias which "act" has and which "proximity" does not have. The Logos system differentiates nouns between nouns that have and do not have a verbal bias.

On the basis of this simple differentiation alone, we resolve the ambiguity for MT vis à vis "revolving" in the following:

(44) a) There are several types of revolving credit.

b) There are several ways of revolving credit.

The well-known, much-discussed linguistic problem raised by the sentences:

(45) a) John is easy to please.

b) John is eager to please.

is sufficiently clarified, at least for the purpose of MT, simply by seeing that the adjective "easy" is verbal, that is, its proper object is a verbal of some kind (e.g., an easy swing, an easy method), whereas "eager" is not intrinsically verbal (e.g., an eager disposition, an eager type). Thus, "eager" modifies John, "easy" modifies "to please". For this reason, therefore, we can say "to please John is easy" but not "to please John is eager".

After verbs themselves, the part of speech that serves most to establish relationships is the preposition. Prepositions are verbal elements, therefore, whether the prepositions are used to complement nouns or verbs. Noun-complementing prepositions. in fact, are verb surrogates:

(46) a) the book on the table

b) the picture on the wall

c) the speaker on the platform

The preposition "on" is a kind of contraction for "lying on" in a), "attached to" in b) "standing (or seated) on" in c).

Verbal elements are no less present in simple noun phrases, albeit in a purely implicit way:

(47) green trees = the trees are green

a milk bottle = a bottle made to hold milk

a glass bottle = a bottle made out of glass

a deposit bottle = a bottle requiring a deposit

These implicit relationships within a compound noun phrase can be ferreted out in The Logos system with some success using simple rules based on combinations of semantico-syntactic codes that are assigned to nouns. This in turn enables the system to explicitate and correctly translate the implied preposition in, e.g., romance languages. In French, the default rendering of the noun compound N1 N2 is N2 de N1. In some cases, as in the case of the English *wine glass*, such a rendering is simply wrong.

In the example of compound nouns (NMN) below, we illustrate how a generalized rule, using semantico-syntactic codes for nouns in the Logos system, can make explicit the implied preposition and thus render it correctly in French.

(48) gold watch: N1(mass material) + N2(functional device) → N2 made of N1

→ montre en or

49. computer tape: N1(agent) + N2(functional mass) → N2 for use by N1

    → bande pour ordinateur

50. wine glass: N1(liquid mass) + N2(container) → N2 à N1

→ verre à vin

The key to a successful MT system lies in mastery over verbal elements (explicit or implicit), which means mastery over their semantico-syntactic properties. The

verb or verbal element orders the relationships of all the other elements in a sentence. (Where there is a relationship there is a verb, explicitly or implicitly.) These relationships are grasped, then, to the extent that the verb element is grasped. Though language (English) has hundreds of thousands of words, the number of verbal relationships are relatively small. There are rougly 10,000 verbs in English which group themselves quite naturally, as we have seen, into a hierarchical arrangement based on their semantico-syntactic properties.

In all, the Logos system grammar has compressed these 10,000 verbs into approximately 100 semantico-syntactic categories at one level of specificity on the semantico-syntactic tree, and into only 15 semantico-syntactic categories at a still more general level. Nouns have about an only modestly larger number of codes on the noun tree at these levels.

Virtually all the linguistic problems described in this document are resolvable by rules employing semantico-syntactic codes at these generalized levels. It is by means of such rules that the human grasp of a sentence is simulated, to the degree needed to translate correctly. That this simulation is possible with relatively such few codes testifies perhaps better than anything else to the correctness of the linguistic principles of Semantico-Syntactics.

FOOTNOTES TO CHAPTER I

1. (page 1) "Semantic equivalence" in this definition does not mean word-for-word translation but does indeed mean literal translation. Literal translation is called for in translation where information transfer is of paramount consideration.

2. (page 1) Manipulations of the type (1) through (3) are generally sufficient to formulate grammatically correct target language sentences. Manipulations (4) and (5) are necessary to produce stylistically correct sentences. The line between (1)-(3) and (4)-(5) tends to separate low-end MT systems from high-end systems

3. (page 5) It appears these abstract elements at the deep structure level are meant to represent semantic thought in some pre-grammatical way, although Chomsky clearly cannot describe them without making use of grammar. In this respect, Chomsky's deep structures resemble Kant's Ding an sich.

4. (page 7) In (11)(iii) both the verb "want" and the syntax of the sentence in which this verb appears express a certain indirectness vis à vis "Mary". These verbs are not convertible, therefore, without muddying the meaning of the sentences as a whole. This suggests, does it not, that the syntactic construction in (11)(iii) and (12)(iv) is different because the verbs are different, and that the syntax reflects the nature of this difference, viz. the degree of transitiveness each of these verbs express, which is itself a function of the verb's semantic value. The degree of transitiveness that verbs exhibit will be seen to have considerable importance in the linguistic principles behind MT.

5. (page 7) The verb "expect" belongs to a different but closely related semantic family of verbs. The lesser degree of directness or transitiveness in "expect" as opposed to "ask" accounts for the fact that the verb "expect" (in French "s'attendre à") falls more naturally into the construction (12)(ii) than does "ask". This may be seen from the opposite fact, namely,.that "expect" and "s'attendre à" do not fit as comfortably in (12)(iii), whereas "ask" and "demander" do. The reason, we suggest, is that the constructions in (12)(iv) and (12)(iii) do not comport with the weaker degree of transitiveness of the verb "expect" and "s'attendre à".

6. (page 7) The Logos system attached special importance to the degree of transitivity in verbs. The following is an abridged description of verbs placed on a scale of transitiveness (see also Figure 1).

a. Intransitives. Verbs that take no direct object are intransitive. Even intransitive verbs, however, very in degree to the subject's relationship to something outside the subject (i.e., in their degree of implied transitivity). For example:

(1) John is at school. (existential intransitive)

(2) John enters the classroom. (motional intransitive)

(3) John listens to the teacher. (operational intransitive)

Intransitive verbs do not necessarily put the subject in a relationship to something else. We can easily say "John is cold", "John runs wild", and "John grows quickly", but the fact is most utterances do in fact establish a relationship between the subject and something else, whether the utterance entails transitive or intransitive verbs.

b. Simple transitives. Verbs that take a single direct object are simple transitive. There are degrees of transitivity even in the case of a simple direct object, ranging from weak to strong.

(1) John studies French. (subjective transitive)

(2) John fights Henry. (reflexive transitive)

(3) John fixes the engine. (objective transitive)

In (1) the subject "John" is the true recipient of the action and, despite its transitive grammatical form, the verb is nearly intransitive. Most languages have a reflexive variant for verbs of this type (e.g., s'appliquer à), including English ("apply oneself to"). In (2) the action is reciprocal, and again the verb is frequently reflexive. In (3) the recipient of the action is the direct object. This is the most purely, most strongly transitive of all verb categories. It represents an exceedingly large class of verbs. Hereafter, the verb categories represent a progressive weakening of transitivity.

c. Simple transitives complemented. Simple complemented transitives required a prepositional phrase to complement the verb. For example: in "John deprived his friend of the opportunity," the prepositional phrase complements the verb and the verb is incomplete without it. There are many verbs complemented in this fashion, using one or another of the prepositions as satellites, so to speak, of their own meaning.

d. Split transitives (di-transitives). Verbs that take two objects, one direct, the second indirect, are split transitive: the immediate recipient of the action is the direct object, the ultimate recipient of the action is the indirect object.

(1) John told Henry a lie.

(2) John gave a book to Henry.

e. Preverbal transitives, simple and split. Verbs whose direct objects are another verb, a nominalized verb or a verbal construction help transitive verbs but are not in themselves fully transitive. We call them preverbals. There are three basic kinds:

(1) John helped repair the engine.

(2) John made the repairs on the engine.

(3) John let Henry repair the engine (authorized Henry to repair the engine).

The so-called object of the verb is another verb in (1), a nominalized verb in (2). The verbs "help" and "make" are called simple preverbals. In (3) we have the introduction of an indirect object which becomes the logical subject of the second verb. Preverbal verbs that allow this more complex complementation we call split preverbal. The construction in (3) borders on the next category, preclausal verbs, where the object of the verb is a clause with a new grammatical subject. Constructions like (3) often are expressible in preclausal form, indicating their close relationship (more broadly and naturally in German than in English).

f. Preclausal transitives. Verbs that introduce a new subject and verb (i.e., a new clause) are preclausal. They are of two basic kinds, (1) simple and (2) split.

(1) John said that Henry made no mistakes.

(2) John informed Henry that he made no mistakes.

In (2) the verb takes an indirect object with the preposition implied. A third preclausal type takes a pseudo direct object and is rather rare in English:

(3) John arranged it that Henry would repair the engine.

7. (page 7) Each part-of-speech has a tree structure of its own.

8. (page 8) See footnote 6 for a discussion of split transitive verbs.

9. (page 9) The interrelating of syntax and semantics occurs within the conventions of a given language. Such conventions are indeed arbitrary, as for example the fact that in simple Persian sentences, the verb comes after the object. But within the framework of these conventions, syntax is sensitive to the

semantic character of the verb. It is true that the individual idiom of each language tends to develop this in ways that are unique to that language, but it is also true that most languages can nearly duplicate the idiom of another language. Though the resulting construction may be strange, it is meaningful and unambiguous, and therefore potentially available to the language. Language conventions are more or less open-ended, the rule being to allow constructions that seem pleasing or useful so long as they avoid ambiguity. One need only look at the headlines of newspapers to appreciate the freedom possible in language provided this rule isn't broken. Moreover, constructions that look very foreign have a way of creeping in when useful. For example, in Russian the participial complement of a noun such as "standing in the corner" is often used adjectivally. Thus: "the standing in the corner man". This sounds utterly foreign to English, but is it? Consider the English expression: "a fun loving, up and coming generation". In English we think that the object of the verb always follows the verb in active declarative sentences. But do they? Consider this construction: "Money I have. It is contentment I seek." In poetry we think nothing of "nor he her love implore 'til she the lie foreswore," exactly as in Persian.

10. (page 9) We must note, parenthetically, that while it appears that the prepositional phrase in (18), "with N3," functions as a verb complement in the manner that we have described, this is by no means an established fact for (18). The syntactic string (18) is rich in ambiguity at this level and capable of several other quite different interpretations. For example:

(19) N1 V1(supply-type) N2 with N3(alacrity-type)

N1 V1(supply-type) N2 Adv.-of-manner

Here, "with N3" functions purely adverbially.

The prepositional phrase "with N3" is purely a noun complement in an (18) having the following semantic values:

(20) John furnished pens with ink.

Despite the strong affinity between the verb and the preposition "with" in this sentence, the complementary relationship between N2 and N3 takes precedence.

The prepositional phrase "with N3" functions as a verbal complement in an (18) having the following semantic values, but does so in a way unrelated to the usual behavior of split transitive verbs such as this.

(21) Manufacturer furnishes ink with pens.

Here the "with" has the special sense of "together with furnishing".

It is obvious that the syntactic string (18), even though the verb has been identified as one making a strong bid for the preposition "with" is fraught with ambiguity that only consideration at the full sentential-semantic level can resolve. The Logos system does this by examining the semantico-syntactic trees for all the elements, considered in light of each other as parts of a total sentence. In doing so, the system employs the natural line of reasoning that the mind takes, as we shall have demonstrated.

11. (page 9) Jean met l'homme au courrant des faits. "Des" has the value of "about the".

12 (page 9) Jeanne ne s'associe pas aux élèves.

13. (page 10) At first glance, the relationship between "with" and "about" in (26) seems forced and arbitrary. Probably the "with" here is actually a contraction of "with (by means of) information about". As for "with" meaning "by means of", this could be a contraction of "together with such and such an instrument" (in the sense that the agent performs the action "together with" the instrument). Certainly prepositions as such are virtually always contractions of a longer expression, a kind of shorthand. "The book on the table" is a contraction for "the book lying on the table". The point of this being that though the preposition undergoes considerable semantic change when preempted by a verb, there must be some reason for the convention which decrees this preposition shall be used as opposed to any other. The most likely reason may lie in the fact that some semantic point of contact exists between the uses of a preposition as a verb complement and as a noun complement, a point of contact that may be buried within a contraction such as we have described. Given this, the gifted student latches onto this additional hint from the preposition itself as to how the preposition is to be rendered in conjunction with certain verbs.

14 (page 11) In the present case, the prepositions fulfill an expectation created by the verb. There are verb-preposition situations, however, where something quite different takes place. For example, a verb like "keep" has several meanings, the proper selection of which depends on some other clue elsewhere in the sentence. Sometimes the clue is a preposition (verb-complementing). For example, in Chomsky's famous sentence: "John kept the care in the garage." the nuance of the verb "kept" is not established until we encounter the prepositional phrase denoting where. Having the "in the garage" as an adverb of place enables us to select the nuance for "kept" correctly. But the process here is rather a reverse of the process taking place between "supply" and "with". The movement is from the preposition back to the verb. What alters is not the nuance of the preposition, but the nuance of the verb itself. Chomsky claims that the sentence "John kept the car in the garage" is actually ambiguous. We do not agree. Anyone who uses that sentence to mean "John kept the car which was in the garage" is viciously violating simple rules of communication. It is not true that "in the garage" could just as easily complement the noun "car", because the verb

"kept" raises a question about itself such that any prepositional phrase of place will be seized upon by the mind as the answer. Another Chomskian example: "John saw Henry walking through the library", we maintain is not ambiguous either, for the simple reason that a complement always complements the thing nearest to it unless a compelling reason exists to override that rule. In the "saw" sentence, no such compelling reason exists as it did in the case of the "keep" sentence. Anyone who would utter such a sentence intending the participial clause "walking through the library" should complement "John" rather than "Henry" doesn't want to communicate.

15. (page 11) Semantico-syntactic codes are labels for what the mind perceives in dealing with language much as the word "chair" is a label for what the mind perceives when it considers a place for the body to sit down. In neither instance, of course, does the label itself spring to mind as the mind deals with the reality for which it stands. It is only in order to communicate about its own operations that the labels themselves become important. Unfortunately, no one hitherto has studied the mind's linguistic operations sufficiently to substantiate this point with respect to those operations. But the semantico-syntactic codes for a given word class are indeed abstractions for words of that word class, much as chair is an abstraction for a place of rest for the body. The mind deals with both at a level quite removed from particular chairs or particular words. For example, when the mind seeks a translation into Russian for the preposition "through", it must translate "through" as "across" in some contexts (thresholds, doors, windows, barriers) and as "along" in others (pipes, conduits). It knows unswervingly to do so not because it has learned which preposition to use for each noun, but because it grasps these nouns abstractly, by this semantico-syntactic property, as it affects the preposition. That is, one class of noun (threshold type nouns) imparts to the preposition "through" the nuance "across; another class of nouns (conduit-type nouns) imparts to the preposition "through" the nuance "along". In the moment that the mind selects the preposition, the mind is not dealing with a particular noun but with an abstraction which, if it were to be labeled, would look like this semantico-syntactic code.

16. (page 14) As he gains proficiency in seeing analogies, the student will very soon relate a new verb not to some other verb having similar semantico-syntactic properties, but to the class itself, the class of all such verbs having this property. In other words, the entity that he likens a new verb to is the semantico-syntactic classification itself (and the conventions which apply to the individual words now apply to the classification). In so doing, the linguistic operation which the student undergoes and the operations which the Logos system perform are identical.

17. (page 15) The point of similarity between the verbs is not purely semantic, strictly speaking, but semantico-syntactic. The student knows that "hit" and "block" are analogous in respect of their use of the proposition "with" to complete their thought; both verb excite an expectation as to means. This property is neither semantic nor syntactic, but semantico-syntactic. The student, therefore, in

encountering a new verb may not explicitly relate the unknown verb to a specific known verb. Instead, and more probably, he relates the new verb to a class of verbs having the same semantico-syntactic property that he feels the new verb must also have. That is, he relates the verb to a known behavior pattern which the inner character of the new verb seems to suggest is appropriate. He knows that other verbs with a similar inner character (semantico-syntactic property) behave thus and so, and feels confident of his ground in using the new verb thus and so. In so doing, the student has actually begun to use a semantico-syntactic shorthand for language. It is just such a shorthand that the Logos system employs.

18. (page 18) Says Miss Sager:

"With regard to syntactic ambiguity, the largest number of cases have their source in different possible groupings of the same string components of the sentence, the decisive criteria being which of the resulting word-associations is the correct one for the given area of discourse. For example, in "changes in cells produced by digitalis" only one of the possible groupings (that in which digitalis produces changes, not cells) is correct within a pharmacology text dealing with cellular effects of digitalis. Recently it has been shown that it is possible to incorporate into the grammar which is used to analyze texts in a given science subject, additional grammatical constraints governing the well- formedness of certain word combinations when they are used in that subfield. These constraints have the force of grammatical rules for discourse within the subfield (not for English as a whole) and have a very strong effect in further informationally structuring the language material in the subfield, and pointing to the correct word association in syntactically ambiguous sentences." ("The Linguistic String Parser", R. Grishman, N. Sager, C. Raze and B. Bookchin, p. 433, Proceedings of the National Computer Conference, 1973)

19. (page 19) Strictly speaking, word association does appear as the only way to resolve ambiguity if a clue is not forthcoming from the verb or verb surrogate. This is not the case in (33), however. As for our word-association example in (35), probably a test could be made based on the verb of the sentence:

(35) Brutus will address the citizens of Rome and friends.

To determine whether "friends" is governed by the preposition "of" or by the verb, it would probably be possible to fashion a rule such that "friends" is a "well-formed" object of the verb "address" and thereby resolve the ambiguity. But the Logos system uses a more elegant method, based on symmetry. All it need ascertain is whether the noun in question is analogous vis à vis semantic type to the first object of the verb, "citizens", and if so to treat the noun in question as another object. The semantic types involved are quite gross: human-type nouns.

Normally where there is no help to be derived from the verbal element, as in (36) and (37):

(36) He carried a silk hat and cane.

(37) He wore a silk hat and tie.

The Logos system falls back on word-combinations but of this very simple kind. If the nouns are homogeneous types as in (37), the modifier of the one is assumed to modify both. If they are alike, the assumption is the opposite. This seems to be much the way the mind works in settling this question for itself. In any event, it is a happy far cry from the notion of "well-formedness of word combinations". Of course, no system can be devised that won't blunder over very elementary intelligence. The system would find it hard on this basis to distinguish syntactically between "happy women and children" and "old women and children". Happily such instances are not too frequent in real-world documents.

II. CONCEPTUAL OVERVIEW OF SEMANTICO-SYNTACTIC OPERATIONS IN THE LOGOS MT SYSTEM

A. General Remarks

The Logos MT system does not deal with an input sentence on the level of the literal word string. Instead, the system converts this sentence into a string of semantico-syntactic codes (expressed as numeric values) and it is on this string that the system operates.

To grasp the nature of these operations, certain basics concepts need to be discussed. These are: (i) semantico-syntactical strings by which the input sentence is represented internally; (ii) linguistic rules consisting of (a) semantico-syntactic patterns which are matchable against the input strings, and (b) an action component which operates on these input strings when a match occurs, e.g. to rewrite the pattern more abstractly, as in parsing. (NB - There can also be context-sensitive constraints, not discussed here, which must be satisfied in a given rule for a match to occur. Constraint satisfaction was not a part of the system at the time this document was written in 1977 but has since come to play a very critical role in rule matching.) To understand the relative power of these linguistic operations, some understanding must also be gained of (iii) levels of specificity in semantico-syntactic codes; (iv) concatenation; (v) normalization.

B. The Semantico-Syntactic String

Let us suppose sentence (1) is read into the Logos System:

(1) John gave his friend several books.

To process (1), the system converts the actual words of (1) into semantico-syntactic codes derived from the semantico-syntactic map for each part of speech. These codes have been associated with these words when the words were originally entered into the system dictionary. Hence the con version from a natural language string to a semantico-syntactic string is a by-product of dictionary look-up.

1. Semantico-Syntactic Codes

Every source-language word entered into the system's dictionary is assigned four semantico-syntactic codes. The lexicographer determines the appropriate set of four codes by locating the word on the appropriate semantico-syntactic tree. There is one such tree for each word class (part-of-speech) in the Logos grammar. These four codes represent four levels of specificity as to semantico-syntactic properties which the word has. In terms of the semantico-syntactic tree, the four levels refer to: (0) the trunk; (1) the limb; (2) the branch; and (3) the sub-branch or twig on which the element is found (see figure 2). We may think of them also as class, superset, set and subset classifications, respectively.

(2) Subset = twig = subset

Set = branch = set

Superset = limb = superset

Class = trunk = word class

For example, the verb "supply" would have the following numeric semantico-syntactic codes:

(3) Level of Numeric Code Semantico-Syntactic

Specificity Property (valence)

3 - Subset 358 V N1 to N2; V N2 N1; V N2 with N1

2 - Set 47 V N1 to N2; V N2 N1

1- Superset 16 V N1 to N2 (split (di-) transitive)

0 - Class 2 Undifferentiated verb

We observe in (3) a progressive narrowing of specification as to the semantico-syntactic properties in the verb "supply". At the lowest or broadest level of specification, Class, "supply" shares the property "verb" with every other verb in the language. At the highest level, Subset, the semantico-syntactic property "supply books to John," "supply John books," "supply John with books" is shared by a tiny cluster of verbs (e.g. supply, furnish, provide).

The four levels of specificity allow rules (patterns) or individual elements of rules to be written at various levels of generality: Thus, rules for performing the various manipulations involved in translation can be made to apply to broad or narrow categories of verbs, as the case requires. How this is accomplished will become abundantly clear when we see the way the linguistic rules operate in The Logos system.

Converting to the Semantico-Syntactic String

The result of the lookup and conversion of (1) to a semantico-syntactic string is as follows. (Not shown are morphological codes for gender, person, number, case, case governance and tense.)

(4) (ss) (ss) (ss) (ss) (ss) (ss) Subset codes

(ss) (ss) (ss) (ss) (ss) (ss) Set codes

(ss) (ss) (ss) (ss) (ss) (ss) Superset codes

N V adj/pro N art N/V Class codes

John gave his friend several books

At the lowest level of specificity, namely Class, the input string (1), now converted to semantico-syntactic (ss) codes in (4), is fraught with ambiguities. To begin with, ambiguity exists in (4) at the lexical-syntactic level because two words in the sentence are syntactically ambiguous: "his" was found in the dictionary to be both a possessive adjective or a possessive pronoun; and the last word in the sentence, "books", was found to be both a verb or a noun. Once these primitive lexical-syntactic ambiguities are resolved, new ambiguities come to light at the sentential-syntactic level, namely, ambiguity as to the relationship between the two noun phrases ("his friend several books") and the verb "gave". Ambiguity at this sentential-syntactic level in turn can only be resolved by resorting to the semantic level, i.e. to clues at the semantic end of the semantico-syntactic spectrum for the words involved. When this is done, the ambiguity is resolved quite readily.

There are a number of different relationships possible at the purely syntactic level in (4), that is, between a verb and two following noun phrases:

(5) John gave his friend a book.

John gave his friend a shove.

John considers his friend a genius.

John bought his friend a dinner.

As we have said, the system resolves these and all such ambiguities by means of linguistic patterns (rules) operating on the semantico-syntactic input string. Another way of stating this is that the string (4) at the Class level (pure syntax) will almost inevitably be manipulated incorrectly unless rules exist which connect with the higher semantico-syntactic properties of (4) and apply the correct interpretations and related manipulations. It is obvious that an ad hoc rule can always be written to handle (4) or any linguistic situation. Just as obviously, unless rules are generalized, unless relatively few rules can handle a great many linguistic situations, the problem of MT cannot be considered as solved.

C. Linguistic Rules

1. Concept of a Generalized Rule in The Logos System

We stated that the Logos system begins its translation of a source language sentence with the conversion of that sentence to a semantico-syntactic string, and that it is this string that undergoes analysis and, in the transfer stage, the manipulations described in Chapter I, which will eventuate in a semantically equivalent target language sentence. The analysis and manipulations on this input string are performed by linguistic rules, the essential characteristics of which we shall now describe.

Linguistic rules in the Logos system are not part of the program fabric of the system. Instead, they exist as data, stored in files, upon which the programs operate. On the other hand, the character of these rules is such that they in turn "drive" the Logos system programs, causing these programs to implement parsing decisions and perform the various manipulations involved in translation.

Linguistic rules in the Logos system consist of two parts, the first concerned with recognizing linguistic situations, the second with effecting the manipulations appropriate to those situations (a third part having to do with constraint satisfaction is not dealt with here, being an innovation to the system subsequent to this writing in 1977). The first part of the total linguistic rule is called an SP rule (for semantico-syntactic pattern). An SP rule consists of a string of semantico-syntactic codes like those used to represent the semantico-syntactic input string.

(NB - This representational homogeneity between the linguistic knowledge base and the input stream has profound computational implications for the system, a topic not dealt with here but which deserves mention because of the contribution this consideration has had to the overall architecture of the Logos model. It is this representational monotonicity that allows the input stream to serve as search argument to the knowledge base, much the way a natural language word is search argument to a dictionary, only here the arguments consist of semantic-syntactic strings. As a result, the size of the rule base has at worst strictly linear impact on performance, much as in the case of dictionaries.)

The semantico-syntactic codes as used by the SP rule, although taken from the same semantico-syntactic trees, usually tend toward the lower levels of specificity. The reason for this should be obvious: the less specific the codes in these rules, the more general their application. Rules are written to be as generalized as possible (which also implies they are written to be as specific as necessary). For example, a single semantico-syntactic rule using a common superset code for the nouns in question allows the system, in each of the following phrases, to parse N Prep N → NP and to recognize the sense of the preposition "on" as "on the subject of" (au sujet de).

(6) books on war

lecture on beauty

report on profitableness

The SP (pattern) portion of the rule that will parse these noun phrases (both syntactically and semantically) looks like this:

(7) N1(superset x) + Prep("on") + N2(~subset y) → NP / (on = "on the subject of")

where superset x is a semantico-syntactic superset code for a broad class of nouns (having to do with information) that govern the preposition "on" and "about" and give these prepositions the meaning of "on the subject of"□; and where subset y denotes surface-bearing type nouns. In the present case, the rule tests that the object of the preposition is <u>not</u> such a noun type.

SP rules tend to be short, for greater applicability, but also need to be as long and as specific as necessary in order to resolve potential ambiguities. For example, if N2 in rule (7) were specified merely by a Class code (denoting any noun whatsoever), clearly the system would mishandle the following noun phrase:

(8) books on shelves

The second portion of the linguistic rule in the Logos system, as we said, is what instructs the Logos system programs as to which manipulations are to be performed. This portion of the linguistic rule is called VTR (for vector transform). Its properties will be discussed elsewhere in this documentation series.

2. Concept of Concatenation in the Logos System

The VTR portion of linguistic rules performs a number of hidden manipulations concerned with the analysis of the semantico-syntactic input string and the resolution of ambiguity, i.e. with the parse.

The principle parsing operating consists of concatenating -- bringing together -- all modifying elements under the heading of the thing they modify, called the head element. Parsing occurs in a compositional manner over a series of separate passes of the sentence. For example, the input sentence:

(9) John bought five interesting new books on the Second World War.

will first be reduced (or concatenated) to its head elements as in (10), illustrated here in natural language, but in reality the concatenated sentence is in semantico-syntactic form:

(10) John bought books on war.

and at a subsequent state will concatenate to simply:

(11) John bought books.

Rule (7), which effects the transformation of the preposition "on", would not match the input string at the stage shown in (9), but does match it in (10), after all simple noun phrases have been concatenated.

The point of interest here is that concatenation, as illustrated in (10), reduces an endless variety of sentences to a single structure suitable for matching on rule (7). By applying these SP pattern rules at the appropriate stage of concatenation, a great economy in the application of these rules is achieved.

The fact that parsing (analysis, concatenation, disambiguation) is conducted compositionally allows the system to pick its way through potentially ambiguous sentences such as (12)

(12) John placed the book on the Second World War on the shelf.

As the parse of (12) unfolds across its various stages, the connominal character of the noun phrase *Second World War* and the converbal character of the noun phrase *on the shelf* are properly recognized. Prepositional phrase attachment is

clearly a critical work of any parser; to guide the Logos system parse in this regard, the VTR portion of linguistic rules, at various stages of analysis, label prepositional phrases as "strong converbal", "weak converbal", "strong connominal", "weak connominal", or adverbial (of varioius types).

3. Concept of Normalization

In certain critical contexts, when a match on an SP rule is made, the VTR of that rule will normalize the matched pattern and resubmit the input string, now normalized, to the knowledge base for further (usually semantic) handling. The benefit of input string normalization is that one SP rule in the knowledge base can cover a great many structural variations on the input side.

The sentences in (13) will illustrate. Because of normalization, one deep-structure rule suffices to disambiguate the meaning of *temper* in all the following sentences.

(13) (a) The wine was destined to be tempered.

(b) Such wine tempering practices are now forbidden.

(c) The tempering of strong wine is an ancient practice

(d) The wine my brother bought was tempered

(e) Wine, once tempered, begins to turn.

This is how it works. At appropriate stages of analysis, when a match on an appropriate SP rule occurs, the VTR of the rule selects the verbal element and its object and presents them in normalized form to a special set of deep structure SP rules form. Collectively these rules form something known as the semantic table. The normalized pattern both of the input string and the rule in the semantic table has the form:

(14) V N

It should be noted that V in this case is more the verb *qua* functional concept than the verb *qua* part of speech, since V will match equally on *tempering* whether occurring as true verb, verbal adjective or as a gerund. It will also match on verb and object regardless of word order or intervening embeddings.

The V in this particular semantic table rule is coded uniquely for "temper" and will match any morphological variant of that verb; N has a semantico-syntactic subset code for liquids. In the input string, V and N, of course, have their full range of

semantico-syntactic codes (i.e. at all four levels of specificity). Now, if any of the input string codes for V and N match the codes for V and N in the semantic table (word order is not relevant here), then the VTR of the semantic rule effects the substitution " temper → "dilute". During target transfer, a similar substitute occurs: tremper → diliuer. Finally, in generation, the morphological form appropriate to the target surface structure is then applied.

D. Conclusions

The point of this discussion has been to show how generalized semantico-syntactic codes, and operations like concatenation and normalization, taken together, permit a linguistic rule as simple in form as "V(type x) N(type y)" to "recognize" and semantically transform a verbal element like "temper" no matter what morphosyntactic context the verb is in, as illustrated in (13). There could hardly be greater elegance than this in devising a linguistic rule for recognizing a particular nuance of the verb "temper". Nor could there be a closer simulation, we think, of the human mind in linguistic operations of this sort, for surely the human mind, when it encounters the verb "temper" in all its variety of usages and syntactic environments, has recourse to a substratum of understanding where the verb's nuance is dealt with conceptually, as a function of its object, independent of other syntactic considerations. In this respect, Chomsky would seem to be correct in postulating the existence of a deep structure in language where surface syntax has no immediate apparent significance. Where Chomsky errs, we believe, is in not perceiving that syntax is latent to semantics and that the seemingly arbitrary caprices of grammatical convention take place, nevertheless, within the limitations of a semantico-syntactic order.

We believe we have grasped this order and mapped it in the various semantico-syntactic trees, to a degree sufficient for purposes of MT, to a degree, that is, sufficient to enable the Logos system to simulate human translation at an acceptable, useful level. No machine will replace a human mind, obviously, but it is clear that machines can simulate some human operations, such as computations, better than the human. Is this also feasible of those operations involved in translation? Certainly not where style is important, when those properties of language are involved whose purpose is to edify rather than inform. But when it comes to information transfer, pure and simple, the very absence of subjectivity and taste in the machine can sometimes be an advantage. While this may be hotly disputed, there has been ample experience in real world situations to bear this out.

It is not claiming too much, however, to predict that a system with a complete set of linguistic codes will always translate with greater consistency of usage and will sometimes translate with greater precision of nuance than the average translator is capable of. The average translator, for example, may not always render the verb "temper" with the precision which the Logos system exemplifies in (15):

(15) John tempered the steel. Jean a trempé l'acier.

John tempered the wine. Jean a coupé le vin.

John tempered his attitude. Jean a adouci son attitude.

Insofar as translation concerns itself solely with the transfer of information between two languages, then the MT system's unique capacity for purely slavish fidelity to the given text, accuracy of nuance and uniformity of usage, all constitute a benefit that should assure such systems a vital role in the world of international commerce, science, technology and law and so forth.